

UNIVERSIDADE FEDERAL DO PARANÁ

ELAINE CONCEIÇÃO VENÂNCIO SANTOS

**MODELO DE MAPA DO CONHECIMENTO EM INSTITUIÇÕES: MINERAÇÃO DE
DADOS E VISUALIZAÇÃO DA INFORMAÇÃO EM UMA APLICAÇÃO WEB**

CURITIBA

2013

ELAINE CONCEIÇÃO VENÂNCIO SANTOS

**MODELO DE MAPA DO CONHECIMENTO EM INSTITUIÇÕES: MINERAÇÃO DE
DADOS E VISUALIZAÇÃO DA INFORMAÇÃO EM UMA APLICAÇÃO WEB**

Dissertação apresentada ao Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação, Área de concentração: Gestão da Informação e do Conhecimento do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Ciência, Gestão e Tecnologia da Informação.

Orientadora: Prof^a Dr^a Denise Fukumi Tsunoda

CURITIBA
2013

UNIVERSIDADE FEDERAL DO PARANÁ. SISTEMA DE BIBLIOTECAS.
CATALOGAÇÃO NA FONTE

Santos, Elaine Conceição Venâncio

Modelo de mapa do conhecimento em instituições: mineração de dados e visualização da informação em uma aplicação web / Elaine Conceição Venâncio Santos. - 2013.

135 f.

Orientadora: Denise Fukumi Tsunoda.

Dissertação (Mestrado) – Universidade Federal do Paraná. Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação, do Setor de Ciências Sociais Aplicadas.

Defesa: Curitiba, 2013

1. Gestão do conhecimento – Universidades e faculdades. 2. Mineração de dados (Computação). 3. Banco de dados – Gerencia. I. Tsunoda, Denise Fukumi, 1972-. II. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação. III. Título.

CDD 658.4038

TERMO DE APROVAÇÃO

Elaine Conceição Venâncio Santos

**“MODELO DE MAPA DO CONHECIMENTO EM INSTITUIÇÕES
MINERAÇÃO DE DADOS E VISUALIZAÇÃO DA INFORMAÇÃO EM UMA
APLICAÇÃO WEB”**

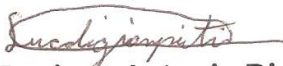
**DISSERTAÇÃO APROVADA COMO REQUISITO PARCIAL PARA
OBTENÇÃO DO GRAU DE MESTRA NO PROGRAMA DE PÓS-
GRADUAÇÃO EM CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO
DA UNIVERSIDADE FEDERAL DO PARANÁ, PELA SEGUINTE BANCA
EXAMINADORA:**



**Prof.^a Dr.^a Denise Fukumi Tsunoda
(Orientadora/UFPR)**



**Prof.^a Dr.^a Helena de Fátima Nunes Silva
(Examinadora/UFPR)**



**Prof. Dr. Luciano Antonio Digiampietri
(Examinador/USP)**

25 de março de 2013

Ao meu esposo Roberto, fonte inesgotável de amor, apoio,
compreensão, estímulo e incentivo.

AGRADECIMENTOS

Agradeço a todas as pessoas que auxiliaram na execução desta pesquisa. De maneira especial, agradeço:

Primeiramente, a Deus e a Maria, por serem meus guias, minha luz e pelas benções que todos os dias concedem a minha vida.

À prof^a Dr^a Denise Tsunoda por sua orientação, ensinamentos, contribuição, apoio, condução e incentivo constante para a execução desta pesquisa. Aos professores membros da banca, Prof^a Dr^a Helena Silva e Prof. Dr. Luciano Digiampietri, por suas importantes contribuições e direcionamentos, ao Prof. Dr. Jesús Mena-Chalco pela valiosa contribuição e observações.

À PRPPG/UFPR, principalmente ao prof. Dr. Sérgio Scheer pelo apoio e incentivo desde a concepção inicial deste projeto, sem o qual este trabalho não teria iniciado, pelas suas contribuições e sugestões.

A todos os professores e colegas do programa que direta ou indiretamente contribuíram para que este trabalho fosse possível, em especial a Ana Carolina e Francisco. Também aos meus amigos de graduação que me acompanham e apoiam até hoje.

À Universidade Federal do Paraná e ao Programa de Pós Graduação em Ciência, Gestão e Tecnologia da Informação da UFPR. À Capes, pelo auxílio financeiro durante o curso.

Aos meus pais por todo amor e dedicação, por quem sou e por tudo que alcancei, por nunca medirem esforços para me ensinar e serem verdadeiros exemplos de vida, de princípios e valores como honestidade e humildade, por me apoiarem em todos os momentos e fazerem sempre mais que o possível por seus filhos.

Aos meus irmãos presentes em todos os momentos como verdadeiros amigos. Meus lindos e amados sobrinhos, que deixam minha vida mais colorida.

Aos meus cunhados e meus sogros sempre carinhosos, atenciosos e Valéria fonte de energia e apoio.

Agradeço imensamente ao meu esposo pelo intenso amor, dedicação, apoio, incentivo, verdadeiro companheirismo, enorme paciência e compreensão, por estímulo e auxílio durante as dificuldades, sem o qual este trabalho não seria realizado e minha vida não estaria completa.

Ao meu primo Jefferson (*in memoriam*) que com certeza estaria comemorando a concretização deste trabalho e obtenção deste título.

Não posso deixar de agradecer também a Yumi, por sua companhia durante todas as etapas da evolução deste trabalho.

Pouco conhecimento faz com que as pessoas se sintam orgulhosas. Muito conhecimento, que se sintam humildes. É assim que as espigas sem grãos erguem desdenhosamente a cabeça para o céu, enquanto que as cheias as baixam para a terra, sua mãe.

Leonardo da Vinci

RESUMO

Este trabalho tem por objetivo propor um modelo para mapear o conhecimento de colaboradores e pesquisadores em instituições. Aborda o problema da perspectiva do compartilhamento do conhecimento científico. O modelo proposto utiliza uma estrutura baseada nos seguintes componentes: atores, grupos de trabalho e artefatos produzidos. O enfoque está na definição do fluxo de atividades necessárias para a construção do mapa do conhecimento e nas alternativas disponíveis para visualização das informações. Com base na revisão da literatura, uma metodologia para apoiar a construção do mapa do conhecimento é apresentada e opções para visualização das informações são propostas. Como prova de conceito, desenvolveu-se um sistema utilizando os dados disponíveis na Plataforma Lattes sobre os pesquisadores e professores dos programas de pós-graduação *stricto sensu* da Universidade Federal do Paraná. A base de dados criada é caracterizada e algoritmos de mineração de dados são aplicados, buscando identificar relacionamentos não conhecidos anteriormente. A validação do modelo proposto é feita por meio do uso da prova de conceito que é composto por dois processos. O primeiro processo é responsável por importar os dados em formato XML para a base de dados do modelo proposto. O segundo processo é responsável por disponibilizar ferramentas para que o usuário possa consultar o mapa do conhecimento e visualizar os resultados utilizando diferentes técnicas como páginas amarelas e grafos. A implementação da prova de conceito auxiliou na identificação dos aspectos positivos do modelo proposto e na detecção de deficiências. Destaca-se como aspectos positivos a criação de um banco de dados especializado, a facilidade de navegação entre as informações disponíveis no banco, a possibilidade de filtrar os resultados das consultas e as diferentes opções de visualização disponíveis, considerando a problemática que norteou a pesquisa sobre construção de mapas do conhecimento que incentivem o compartilhamento e facilitem a localização do conhecimento.

Palavras-chave: Mapa do conhecimento. Visualização do conhecimento. Gestão do conhecimento. Mineração de Dados.

ABSTRACT

This work aims to propose a model to map the knowledge of employees and researchers in institutions. It approaches the problem from the perspective of sharing scientific knowledge. The proposed model uses a framework based on the following components: actors, working groups and artifacts produced. The focus is to define the flow of activities necessary for the construction of the knowledge map and the alternatives available for viewing information. Based on the literature review, a methodology to support the construction of the knowledge map is presented and options for viewing information are proposed. As proof of concept, we developed a system using the data available in the Lattes Platform for researchers and teachers of the *stricto sensu* programs at the Federal University of Paraná. The database created is characterized and data mining algorithms are applied in order to identify relationships not previously known. The model validation is done by the use of proof of concept that consists of two processes. The first process is responsible for importing data in XML format for the database model. The second process is responsible for providing tools so that the user can query the knowledge map and visualize the results using different techniques such as yellow pages and graphs. The implementation of the proof of concept helped in identifying the positive aspects of the proposed model and the detection of deficiencies. Stands out as positives the creation of a specialized database, ease navigation between the content of the database, possibility to filter the results of queries and different display options, considering the problems that guided research on construction of knowledge maps that encourage and facilitate the sharing of knowledge location.

Keywords: Knowledge map. Knowledge visualization. Knowledge management. Data mining.

LISTA DE FIGURAS

FIGURA 1 – CICLO DA GESTÃO DA INFORMAÇÃO	23
FIGURA 2 - ESPIRAL DO CONHECIMENTO	28
FIGURA 3 - COMPONDO UM MAPA DO CONHECIMENTO	31
FIGURA 4 – PROCESSO DE MAPEAMENTO DO CONHECIMENTO	31
FIGURA 5 – FERRAMENTA DE MAPA DO CONHECIMENTO EXIBINDO O RESULTADO DA BUSCA.....	34
FIGURA 6 – FERRAMENTA DE MAPA DO CONHECIMENTO EXIBINDO A OPÇÃO DE VISUALIZAÇÃO DA INFORMAÇÃO DO MAPA “MAP”	34
FIGURA 7 – RELATÓRIO DE PESSOA-CHAVE	35
FIGURA 8 - LEIAUTE DE ÁRVORE MODERADAMENTE GRANDE.....	42
FIGURA 9 - TÉCNICAS HIERARQUICAS DE VISUALIZAÇÃO (A) <i>CONE TREE</i> E (B) <i>CAM TREE</i>	43
FIGURA 10 – ESQUEMA BÁSICO DA TÉCNICA <i>TREEMAPS</i>	43
FIGURA 11 - TÉCNICA <i>CUSHION TREEMAPS</i>	44
FIGURA 12 - TÉCNICA <i>INFORMATION SLICES</i> COM SEMICÍRCULO AUXILIAR PARA APRESENTAR NÍVEIS COM DETALHAMENTO MAIOR	45
FIGURA 13 – UM GRAFO $G(V,A)$	46
FIGURA 14 – EXEMPLO DE UM GRAFO COM VÉRTICES ISOLADOS.....	46
FIGURA 15 – TIPOS DE ARESTAS NO GRAFO	47
FIGURA 16 - VISÃO GERAL DOS PASSOS QUE COMPOEM O PROCESSO DE KDD.....	54
FIGURA 17 – EXEMPLO DO RELATÓRIO DE PRODUÇÃO BILIOGRÁFICA, TÉCNICA E ARTÍSTICA OBTIDA COM O SCRIPTLATTES.....	61
FIGURA 18 – EXEMPLO DE PARTE DO GRAFO DE COLABORAÇÃO GERADO COM O SCRIPTLATTES	62
FIGURA 19 – OPÇÃO PARA BAIXAR CURRÍCULO LATTES EM ARQUIVO FORMATO XML	67
FIGURA 20 - MODELO ENTIDADE-RELACIONAMENTO	77
FIGURA 21 – TELA INICIAL DA PROVA DE CONCEITO COM LISTA DE ATORES POR NÚMERO DE IDENTIFICADOR	97
FIGURA 22 – TELA COM INFORMAÇÕES DETALHADAS SOBRE UM DETERMINADO ATOR	98
FIGURA 23 – TELA DE VISUALIZAÇÃO DO MAPA UTILIZANDO PESQUISA.....	98
FIGURA 24 – RESULTADO DA CONSULTA POR ÁREA DO CONHECIMENTO	99
FIGURA 25 – PESQUISA POR PALAVRA CHAVE	99

FIGURA 26 – RESULTADO DA PESQUISA POR PALAVRA CHAVE.....	100
FIGURA 27 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS	100
FIGURA 28 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS	101
FIGURA 29 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS	102

LISTA DE QUADROS

QUADRO 1 - COMPARAÇÃO DAS TÉCNICAS DE MAPA DO CONHECIMENTO ...	36
QUADRO 2 - CLASSES DE REPRESENTAÇÕES VISUAIS.....	39
QUADRO 3 – DESCRIÇÃO DOS RELACIONAMENTOS PELA CARDINALIDADE ..	51
QUADRO 4 – CARACTERÍSTICAS DOS TRABALHOS CO-RELATOS.....	63
QUADRO 5 - COMPARAÇÃO ENTRE OS MODELOS DA LITERATURA E O MODELO PROPOSTO.....	72
QUADRO 6 - DESCRIÇÃO DAS TABELAS DO MER.....	77
QUADRO 7 - FORMAS DE CADASTRO DE UMA MESMA AREA NA BASE	80
QUADRO 8 - REGRAS DE ASSOCIAÇÃO ENCONTRADAS COM O ALGORITMO APRIORI.....	88
QUADRO 9 - MATRIZ PARCIAL DE CONFUSÃO DO CLASSIFICADOR OBTIDO PELO ALGORITMO C4.5	89
QUADRO 10 - PRINCIPAIS REGRAS DE CLASSIFICAÇÃO GERADA PELO C4.5 ..	90
QUADRO 11 - MATRIZ PARCIAL DE CONFUSÃO DO CLASSIFICADOR OBTIDO PELO ALGORITMO PRISM	91
QUADRO 12 – PRIMEIRAS REGRAS DE CLASSIFICAÇÃO GERADA PELO PRISM	92

LISTA DE GRÁFICOS

GRÁFICO 1 – PESQUISADORES EM UMA MESMA ÁREA DE ATUAÇÃO CONSIDERANDO A ESPECIALIDADE	81
GRÁFICO 2 – ÁREAS DE ATUAÇÃO X PESQUISADORES.....	82
GRÁFICO 3 – QUANTIDADE DE ARTEFATOS POR TIPO.....	83
GRÁFICO 4 – RELAÇÃO DE ARTEFATOS X QUANTIDADE DE AUTOR(ES)	83

LISTA DE ABREVIATURAS E SIGLAS

BD	- Banco de Dados
CAPES	- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	- Conselho Nacional de Desenvolvimento Científico e Tecnológico
CE	- Conjunto de Entidades
DDM	- <i>Distributed Data Mining</i>
DOI	- <i>Digital Object Identifier</i>
GC	- Gestão do Conhecimento
GI	- Gestão da Informação
ISSN	- <i>International Standard Serial Number</i>
JIT	- <i>JavaScript InfoVis Toolkit</i>
KDD	- <i>Knowledge Discovery in Data Base</i>
MC	- Mapa do Conhecimento
MD	- Mineração de Dados
MER	- Modelo Entidade-Relacionamento
PRPPG	- Pró-reitoria de Pesquisa e Pós-graduação
SGBD	- Sistema Gerenciador de Banco de Dados
UFPR	- Universidade Federal do Paraná
UML	- <i>Unified Modeling Language</i>
VC	- Visualização do Conhecimento
XML	- <i>Extensible Markup Language</i>
USP	- Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO	17
1.1 PROBLEMATIZAÇÃO	17
1.2 OBJETIVOS	19
1.3 JUSTIFICATIVA	19
1.4 ESTRUTURA DA DISSERTAÇÃO	20
2 FUNDAMENTAÇÃO TEÓRICA	21
2.1 GESTÃO DO CONHECIMENTO	21
2.1.1 Informação e conhecimento	22
2.1.2 Compartilhamento e codificação do conhecimento	27
2.1.2.1 Mapeamento do conhecimento	30
2.1.2.2 Técnicas de mapa do conhecimento	36
2.2 VISUALIZAÇÃO	38
2.2.1 Representação visual	39
2.2.2 Categorização das técnicas de visualização	40
2.2.3 Técnicas de visualização hierárquicas	41
2.2.4 Conceito de grafos	45
2.3 BANCO DE DADOS	47
2.3.1 Modelos de dados	49
2.3.2 Modelo entidade-relacionamento	51
2.3.3 Relacionamentos	51
2.3.4 Normalização de dados relacionais	52
2.4 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS	53
2.4.1 Processo de KDD	54
2.4.2 Mineração de dados	55
2.4.3 Classificação	56
2.4.3.1 C4.5	56
2.4.3.2 PRISM	57
2.4.1 Associação – Algoritmo Apriori	58
2.5 TRABALHOS RELACIONADOS	60
3 METODOLOGIA DE PESQUISA	64
3.1 O AMBIENTE DE PESQUISA	64
3.2 COLETA DOS DADOS	65

3.3 UNIVERSO E AMOSTRAGEM	67
3.4 MATERIAIS	68
3.4.1 Plataforma Lattes	68
3.4.2 <i>Extensible Markup Language</i> - XML	68
3.4.3 PostgreSQL.....	69
3.4.4 Python	69
3.4.5 Django	70
3.4.6 Java.....	70
3.4.7 Xstream	70
3.4.8 JavaScript InfoVis Toolkit – JIT	71
3.4.9 Microsoft Office Excel.....	71
3.4.10 Weka	71
4 CONSTRUÇÃO DO MODELO.....	72
4.1 COMPONENTES	73
4.2 PROCESSOS.....	75
4.3 PRODUTOS	75
4.4 MODELO ENTIDADE-RELACIONAMENTO DO BANCO DE DADOS	76
4.5 CARACTERIZAÇÃO DO BANCO DE DADOS.....	79
4.6 MINERAÇÃO DOS DADOS DA BASE LATTES	84
4.6.1 Limpeza dos dados e pré-processamento.....	85
4.6.2 Resultado do algoritmo de associação: Apriori	87
4.6.3 Resultado do algoritmo de árvore de decisão: C4.5.....	89
4.6.4 Resultado do algoritmo de regras de classificação: PRISM.....	91
5 APLICAÇÃO DO MODELO	93
5.1 PROCESSO DE CONSTRUÇÃO DA PROVA DE CONCEITO	93
5.2 INTERFACE DA PROVA DE CONCEITO.....	97
5.3 ANÁLISE DOS RESULTADOS	102
6 CONSIDERAÇÕES FINAIS	105
6.1 ATENDIMENTO AOS OBJETIVOS PROPOSTOS	106
6.2 SUGESTÕES PARA TRABALHOS FUTUROS	107
REFERÊNCIAS.....	109
APÊNDICE A – PADRONIZAÇÃO DO DEPARTAMENTO PARA A MINERAÇÃO DE DADOS	117
APÊNDICE B – PADRONIZAÇÃO DA FORMAÇÃO PARA A MINERAÇÃO DE DADOS	119

APÊNDICE C – RESULTADO ALGORITMO C4.5 DAS REGRAS COM INSTÂNCIAS CLASSIFICADAS	129
ANEXO A – PARECER DO COMITÊ DE ÉTICA EM PESQUISA DA UFPR	131
ANEXO B – AUTORIZAÇÃO DA PRPPG PARA COLETA DOS DADOS.....	135

1 INTRODUÇÃO

A atividade de pesquisa e ensino no ambiente acadêmico gera a troca de conhecimento entre diversos atores e envolve pesquisadores/professores e pesquisadores/alunos. Essa troca se dá em diferentes momentos e por meio de instrumentos diversos tais como: exposição oral em sala de aula, seminários, grupos de estudos, projetos de pesquisa ou publicação de livros e artigos. Encontrar um pesquisador, professor ou aluno que trabalhe com determinado tipo de conhecimento na instituição nem sempre é uma tarefa fácil, um dos motivos pode ser atribuído à ausência de ferramentas que permitam localizá-los com base nas áreas de interesse ou trabalhos publicados. A elaboração de um modelo que permita mapear o conhecimento existente em instituições e forneça múltiplas alternativas de visualização dos atores e das áreas em que atuam é o tema abordado neste trabalho.

1.1 PROBLEMATIZAÇÃO

A gestão do conhecimento (GC) apresenta o conhecimento como um agente presente na organização, tanto na mente das pessoas que o possuem ou o adquirem pela experiência, chamado de conhecimento tácito, quanto por meio dos registros existentes, conhecimento explícito (TURBAN; MCLEAN; WETHERBE, 2004).

No que se refere ao conhecimento explícito, é imprescindível que este seja facilmente localizável, caso contrário não agregará o devido valor às organizações que almejam gerenciar o conhecimento institucional. Alguns fatores que dificultam o acesso ao conhecimento já registrado estão relacionados:

- a) a grande quantidade de dados e informações disponíveis na organização de forma não organizada;
- b) a problemas relativos à gestão de pessoas, criando barreiras ao fluxo e divulgação da informação;

- c) a ausência de ferramentas voltadas para a localização do conhecimento já registrado.

Este último fator se embasa na pesquisa de Sasieta (2011), a qual evidencia que embora as representações visuais venham sendo utilizadas como suporte para a disseminação do conhecimento, não há, em contrapartida, propostas de modelos para recuperar e visualizar o conhecimento.

Um desafio às organizações é a adoção de práticas gerenciais que promovam e incentivem o compartilhamento do conhecimento de seus colaboradores, a fim de que o mesmo possa passar do individual ao coletivo, possibilitando, ao menos em parte, ser mantido na instituição, mesmo com a rotatividade dos recursos humanos. Para isso, diversas práticas são utilizadas (como: portais corporativos, reuniões, narrativas, grupos de discussão, bases de conhecimento), visando promover a transferência de conhecimento entre os colaboradores e para a memória da organização.

É necessário ressaltar que se a organização não sabe qual conhecimento seus colaboradores possuem, torna-se inevitável o prejuízo com a perda informacional e com o retrabalho necessário a sua nova obtenção. Nesse contexto, o mapa do conhecimento (MC) objetiva:

- a) localizar e apontar o conhecimento dentro da organização;
- b) mapear os detentores do conhecimento que ali atuam;
- c) a identificação das barreiras que impedem o livre fluxo do conhecimento e das oportunidades que podem alavancar o conhecimento na organização (DAVENPORT; PRUSAK, 1998).

Neste contexto, a questão desta pesquisa consiste em: considerando a possibilidade de sistematizar e classificar as informações dos atores da instituição e sua produção técnica/científica de forma que consultas e visualizações estejam disponíveis em múltiplos formatos, como construir um mapa do conhecimento que facilite o seu compartilhamento e localização?

1.2 OBJETIVOS

Como objetivo geral definiu-se desenvolver um modelo para a construção de mapa do conhecimento para instituições acadêmicas.

Do objetivo geral derivaram-se os objetivos específicos:

- a) construir uma base de dados para armazenamento dos dados referente ao conhecimento de professores e pesquisadores e suas respectivas produções científicas;
- b) identificar relacionamentos e associações nesta base de dados;
- c) validar o método proposto, por meio da implementação de uma ferramenta para prova de conceito.

1.3 JUSTIFICATIVA

A presente pesquisa está associada às práticas de gestão do conhecimento (GC), em específico o seu compartilhamento, como uma proposta que está alinhada à estratégia institucional. Justifica-se pela contribuição que dará na sistematização para compartilhamento do conhecimento, relacionando a literatura com a presente proposta de modelo para mapeamento de conhecimento a ser desenvolvido na Universidade Federal do Paraná - UFPR. Este estudo identifica o processo de compartilhamento de conhecimento existente na instituição e permite localizar possíveis lacunas e/ou barreiras.

Quanto à contribuição da pesquisa para a área identifica-se que os mapas do conhecimento representam uma eficiente ferramenta para auxiliar os profissionais que trabalham com a informação, de forma direta ou indireta. Esta prática de organização do conhecimento proporciona ao pesquisador maneiras objetivas de visualizar as informações.

A pesquisa também auxilia na organização de recursos bibliográficos e na gestão de recursos humanos por permitir identificar os perfis dos pesquisadores, não apenas por meio dos dados que constam em seus currículos, mas também pelas relações identificadas entre as áreas em que atuam. Também há a contribuição para o compartilhamento do conhecimento na instituição, o que vai ao encontro da

política de qualificação e capacitação dos professores/pesquisadores de instituições como da UFPR, por apresentar uma forma de compartilhar o conhecimento entre as pessoas que ali atuam, pode-se assim apoiar a qualificação/capacitação, por meio da identificação dos professores/pesquisadores e dos documentos produzidos por estes.

A relevância acadêmica deste estudo está no desenvolvimento de um método para facilitar o compartilhamento do conhecimento por meio da construção de mapas, possibilitando assim contribuir na análise de outros estudos relacionados às práticas de gestão do conhecimento.

1.4 ESTRUTURA DA DISSERTAÇÃO

O presente trabalho estrutura-se em seis seções e demais complementos. Nesta primeira seção são apresentados: a introdução ao tema, o problema da pesquisa, o objetivo geral, os objetivos específicos e a justificativa do estudo.

A segunda seção apresenta a fundamentação teórica, com a finalidade de embasar as ideias e oferecer aporte para análise e interpretação dos conceitos e dados coletados e discutidos neste estudo. Também são destacados trabalhos similares publicados nos últimos dez anos que contribuíram com esta pesquisa.

A terceira seção discorre sobre a metodologia aplicada para alcançar os objetivos desta pesquisa, apresentando a base utilizada, suas características e limitações e resultados obtidos por meio da mineração de dados, bem como os materiais utilizados.

A quarta seção envolve a construção do modelo de mapa do conhecimento, suas características, o modelo proposto para o banco de dados e descreve a construção da prova de conceito.

A quinta seção apresenta os resultados obtidos com o modelo proposto por meio da prova de conceito realizada e apresenta as formas de visualização propostas para o mapa. A sexta seção é dedicada às considerações finais apresentando vantagens e desvantagens do modelo e as sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os principais conceitos que norteiam este trabalho, baseados em fontes bibliográficas pertinentes para o tema de pesquisa. Inicialmente com fundamentos da gestão do conhecimento, diferenciação de conhecimento e informação, a atividade de compartilhamento do conhecimento e em especial a técnica para mapeamento do conhecimento. A seção de visualização aborda técnicas para representar o mapa do conhecimento. Já as seções referentes a: banco de dados e descoberta de conhecimento em base de dados, conceituam ferramentas utilizadas na construção do modelo. Finaliza com a descrição dos trabalhos relacionados.

2.1 GESTÃO DO CONHECIMENTO

Conforme Turban, Mclean e Wetherbe (2004, p. 326), a GC é um “processo que ajuda as empresas a identificar, selecionar, organizar, distribuir e transferir informação e conhecimento especializado que fazem parte da memória da empresa e que normalmente existem dentro delas de forma não-estruturada”. Comenta também que a estruturação do conhecimento permite a resolução de problemas de forma eficaz e eficiente. Neste sentido, a GC é utilizada com o objetivo de fornecer a informação para as pessoas certas e no momento certo.

Na seção 2.1.1 serão apresentados conceitos para diferenciar informação e conhecimento, além de gestão do conhecimento e gestão da informação. Estas diferenciações são primordiais para a análise das demais seções e para a compreensão da dificuldade de compartilhar conhecimento.

2.1.1 Informação e conhecimento

Como o objeto desta pesquisa não é estudar a epistemologia das definições de informação e conhecimento, as definições que serão apresentadas a seguir são dos principais autores da área.

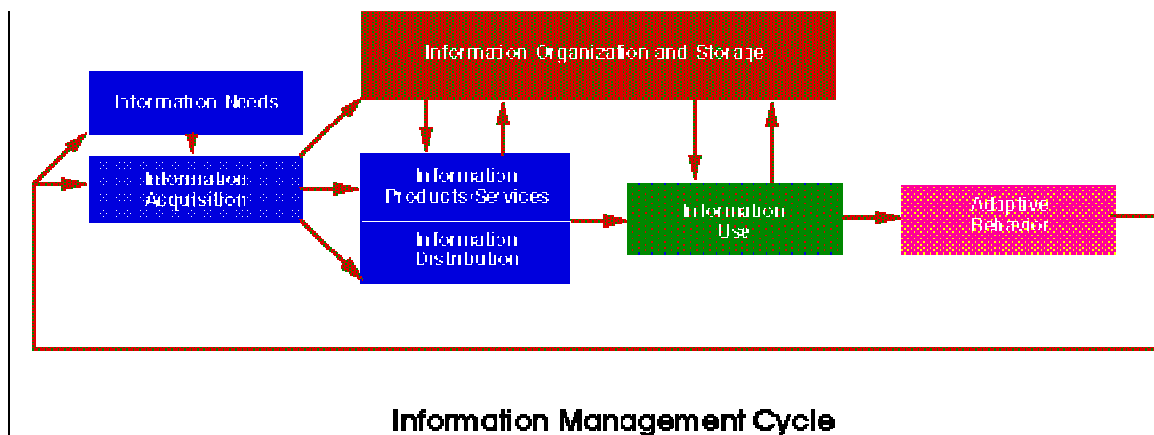
Para Davenport (1998, p. 18-19), a informação é difícil de ser definida, pois envolve dados, a própria informação e conhecimento. Ela serve de conexão entre dados brutos e o conhecimento que se pode eventualmente obter deles. Enquanto define a informação como “dados dotados de relevância e propósito” os dados seriam “simples observações sobre o estado do mundo”.

Os “dados tornam-se informação quando o seu criador lhes acrescenta significado”. A informação é como uma mensagem em que são necessários um emissor e de um receptor e tem “por finalidade mudar o modo como o destinatário vê algo, exercer algum impacto sobre o seu julgamento e comportamento. Ela deve informar; são os dados que fazem a diferença” (DAVENPORT; PRUSAK, 1998, p. 6).

Para McGee e Prusak (1994, p. 5), a informação é dinâmica, cria valor significativo para as organizações, possibilita que novos produtos e serviços sejam criados e aperfeiçoa a qualidade do processo decisório na organização.

A gestão da informação (GI) é um ciclo contínuo de adquirir, organizar, controlar, disseminar e utilizar a informação (MCGEE; PRUSAK, 1994, p. 108). Para Choo (1995) o objetivo da GI é o aproveitamento de recursos da informação da organização a fim de permitir que a organização aprenda e se adapte ao seu ambiente de mudança. A Figura 1 apresenta o modelo onde seis processos de gestão distintos podem ser identificados, além do comportamento adaptativo (*Adaptive Behavior*) que está relacionado à inteligência organizacional. Para o autor, uma organização inteligente persegue seus objetivos em um ambiente externo de constante mudança e adapta seu comportamento de acordo com o cenário interno e externo, sendo caracterizado como um ciclo contínuo.

FIGURA 1 – CICLO DA GESTÃO DA INFORMAÇÃO



FONTE: Choo (1995)

Os seis processos do ciclo de gestão da informação apontados por Choo (1995) são:

- a) requisitos de informação (*Information Needs*): consiste em identificar e representar de forma detalhada e completa a necessidade de informação dos usuários;
- b) aquisição da informação (*Information Acquisition*): é visto como uma função crítica e cada vez mais complexa da gestão da informação. A aquisição da informação busca um equilíbrio entre duas demandas contrárias. A primeira, está relacionada com a necessidade das instituições de uma quantidade crescente de informações sobre os temas de interesse ou relacionados. A segunda demanda está relacionada com a capacidade das instituições de absorver informações, que está ligada diretamente a capacidade limitada do ser humano de absorção, gerando a necessidade de filtrar quais informações ou tipos de informações devem ser filtrados;
- c) organização e armazenamento das informações (*Information Organization and Storage*): este processo pode ser facilitado pelo uso da tecnologia da informação para auxiliar na recuperação da informação. Atualmente, com o grande volume de dados coletados com o uso da tecnologia da informação, tornou-se necessário desenvolver ferramentas que permitam aos gestores visualizar e manipular estes volumes de dados, buscando por padrões ou informações;

- d) produtos e serviços de informação (*Information Products and Services*): tem por objetivo não apenas fornecer informações relevantes para as áreas de interesse do usuário, mas fornecê-las de maneira mais acessível. Desta forma, precisam entregar e apresentar informações que em seu conteúdo, formato e demais atributos atendam aos requisitos situacionais que afetam a resolução do problema ou da classe de problemas. Os serviços de informação precisam ser constantemente inovados buscando satisfazer as diferentes necessidades dos usuários;
- e) distribuição da informação (*Information Distribution*): seu propósito é encorajar o compartilhamento, pois a distribuição da informação promove a aprendizagem mais generalizada e mais frequente. Um meio para a distribuição e a partilha da informação é a utilização de grupos de discussão formados por usuários motivados, ou seja, usuários interessados em compartilhar informações. Este meio tem a capacidade de adicionar profundidade considerável para a base de conhecimento da organização;
- f) uso da informação (*Information Use*): é um processo dinâmico e socialmente interativo da pesquisa que pode resultar na criação de significado ou na tomada de decisão. O uso da informação para a criação de significado e compreensão exige processos e métodos informacionais que ofereçam alto grau de flexibilidade na representação da informação, isto facilita o intercâmbio e a avaliação de múltiplas representações entre os indivíduos. A informação precisa ser compartilhada com facilidade, mas sem perder a riqueza cognitiva. Por meio da troca e interpretação de informações, a organização pode combinar o conhecimento tácito e explícito para extrair novos sentidos das ações.

Choo (2006, p. 362) identifica que as organizações processam e usam a informação em três camadas, sendo elas: “criação de significado, construção do conhecimento e tomada de decisões”. O conhecimento organizacional surge quando as três formas de usar a informação se conectam para construir uma rede maior de processos que continuamente geram significado, aprendizado e ações.

O conhecimento é tido como uma informação valiosa porque alguém atribuiu à informação um contexto, um significado e uma interpretação (DAVENPORT, 1998, p. 18-19).

O conhecimento pode ser dividido em quatro tipos: popular, filosófico, religioso e científico. Sendo que existe uma correlação entre o conhecimento popular e científico, o que os diferencia é a forma, modo ou o método e os instrumentos do “conhecer” (MARCONI; LAKATOS, 2007).

Enquanto o conhecimento popular ou vulgar caracteriza-se predominantemente como superficial, sensitivo, subjetivo, assistemático e acrítico, o conhecimento científico é real (ANDER-EGG, 1978, *apud* MARCONI; LAKATOS, 2007). Para Trujillo (1974, *apud* MARCONI; LAKATOS, 2007, p. 20), este se constitui como contingente, pois sua veracidade é comprovada ou não. É sistemático, por se tratar de um conhecimento ordenado. Possui verificabilidade pela possível comprovação de hipóteses. É falível em virtude de não ser definitivo absoluto ou final, e por isto é exato, diante da condição de reformulação da teoria por ocasião do desenvolvimento de novas técnicas.

Para Davenport e Prusak (1998, p. 6), o conhecimento é frequentemente tácito e de difícil transferência, reúne uma combinação da experiência, de valores, de informação contextual e ideias experimentadas que proporcionam “uma estrutura para a avaliação e a incorporação de novas experiências e informações”.

O conhecimento tácito, conforme Takeuchi e Nonaka (2009), é altamente pessoal e de difícil formalização, o que torna a comunicação e o compartilhamento difíceis, tendo em vista que ele não é facilmente explicável, pois está enraizado nas ações e na experiência do indivíduo, assim como nas ideias, valores ou emoções. Os autores também apontam duas dimensões para o conhecimento tácito: a primeira, “técnica”, abrange as habilidades informais, muitas vezes captadas no termo “*know-how*”; a segunda, “cognitiva”, consiste nas “crenças, percepções, ideais, valores, emoções e modelos mentais tão inseridos em nós que os consideramos naturais”. Quanto ao conhecimento explícito, o mesmo pode ser expresso em “palavras, números ou sons e compartilhado na forma de dados, fórmulas científicas, recursos visuais, fitas de áudio, especificações de produtos ou manuais”, possibilitando que seja rapidamente transmitido aos indivíduos, formal e sistematicamente (TAKEUCHI; NONAKA 2008, p. 19).

Para as organizações, o conhecimento é entendido como a experiência acumulada pelos colaboradores que ali atuam. Choo (2006, p.356-358) resume em cinco as características do conhecimento organizacional, conforme as definições a seguir:

- a) mediado: o conhecimento organizacional é resultado da integração entre os elementos: sistemas de atividades constituídos de indivíduos, rotinas e objetos da atividade. As interações são mediadas por regras formais e informais, uso de ferramentas e tecnologias e pelos papéis e relacionamentos;
- b) situado: localiza-se no tempo e espaço e interage com elementos físicos e sociais do ambiente em que a atividade ocorre. Depende da forma com que as pessoas interpretam o contexto em que atuam, inclusive na percepção e sentimentos em relação à comunidade a que pertencem e na qual desenvolvem suas habilidades;
- c) provisório: novos conhecimentos são testados à medida que as hipóteses e teorias são construídas, experimentadas e reestruturadas. As rotinas, regras e os papéis são constantemente revistos e reconfigurados;
- d) pragmático: produz ação orientada para objetivos, sendo a ação coletiva guiada pelas compreensões que as pessoas têm do objeto de suas atividades;
- e) contestável: devido a muitas vezes ser usado como recurso de poder, o conhecimento, nas organizações, acaba ganhando tons políticos e os diferentes níveis ou restrições de acesso ao mesmo podem criar situações de conflito.

A necessidade de integrar as ações relacionadas ao conhecimento nas organizações é tratada pela gestão do conhecimento (GC). Algumas empresas podem já ter iniciado este gerenciamento sem identificar que o fazem, conforme Davenport e Prusak (1999) observam:

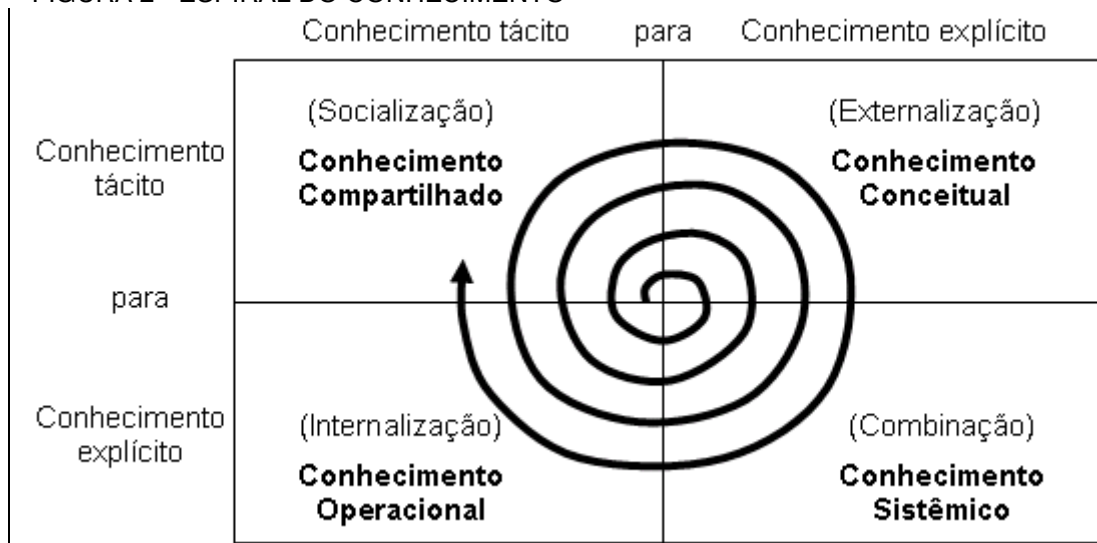
A gestão do conhecimento baseia-se em recursos existentes, com os quais a sua organização pode já estar contando – uma boa gestão de sistemas de informação, uma gestão de mudança organizacional e boas práticas da gestão de recursos humanos. Se você tem uma boa biblioteca, um sistema de banco de dados textuais ou até mesmo programas educativos eficazes, provavelmente a sua empresa já está fazendo alguma coisa que poderia ser chamado de gestão do conhecimento (DAVENPORT; PRUSAK, 1999, p.196).

Um dos fatores que envolvem a GC é o compartilhamento do conhecimento, e sua prática torna-se necessária uma vez que não há sentido em deter conhecimentos importantes se não há promoção de partilha. A organização somente se beneficia como um todo quando o conhecimento é compartilhado e alavancado. Na GC, portanto, um dos grandes desafios é promover o compartilhamento do conhecimento. Devido à dificuldade de captação, formalização e comunicação, o compartilhamento pode necessitar de mudanças e mobilização de toda a instituição. Por isto, compartilhar o conhecimento, a fim de que ele não fique concentrado em algumas pessoas, apresenta-se como um desafio (SATO, 2010, p. 35).

2.1.2 Compartilhamento e codificação do conhecimento

Conforme Sveiby (1998), o compartilhamento do conhecimento é a principal atividade nas organizações e, como já visto anteriormente, ele é de dois tipos: o explícito, que é sistemático e pode ser comunicado e partilhado, e o tácito, altamente pessoal e de difícil formalização. As empresas podem criar conhecimento por meio da interação entre o conhecimento tácito e o explícito – um processo chamado de conversão do conhecimento, havendo quatro modos de conversão: socialização, externalização, combinação e internalização (TAKEUCHI; NONAKA, 2008). A Figura 2 ilustra a espiral do conhecimento.

FIGURA 2 - ESPIRAL DO CONHECIMENTO



FONTE: Nonaka e Takeuchi (1997)

Por meio da conversão do conhecimento também é possível identificar os modos de compartilhamento definidos por Nonaka e Takeuchi (1997):

- a) socialização: é a partilha do conhecimento de tácito para tácito, é um processo de troca de experiências aonde a pessoa compartilha seu conhecimento tácito diretamente a outra;
- b) externalização: é a conversão do conhecimento de tácito para explícito, por meio da articulação para que outros possam compartilhar o conhecimento;
- c) combinação: é a partilha de um conhecimento já explícito de forma também explícita, sendo que as bases de dados podem auxiliar neste processo;
- d) internalização: é o processo de incorporar o conhecimento explícito ao tácito estando estritamente relacionado com o “aprender fazendo”.

Segundo Grotto (2003), por meio destes quatro modos é que se obtém o compartilhamento do conhecimento. No entanto, esta tarefa é simplificada quando o conhecimento é explícito; quando se trata do tácito, revela-se, muitas vezes, demorado e ineficaz. Davenport e Prusak (1998) apontam que a relativa dificuldade de absorver e transferir o conhecimento depende do tipo do conhecimento envolvido. Enquanto o conhecimento mais ou menos explícito pode ser embutido em procedimentos ou representado em documentos e bancos de dados e transferido

com razoável acurácia, a transferência do conhecimento tácito normalmente exige intenso contato pessoal.

Para Grotto (2003), “o compartilhamento do conhecimento pode ocorrer por meio de práticas informais ou formais”. O compartilhamento informal ocorre de forma não preestabelecida durante encontros casuais e conversas, por exemplo. Por ser uma prática informal não está documentado. Sua viabilidade, portanto, depende de conversas e contatos que podem não ocorrer; assim, é importante contar também com práticas formais. No compartilhamento formal utilizam-se algumas práticas como palestras, apresentações, manuais e livros, além dos conhecimentos codificados e armazenados na organização. A pesquisa de Hansen, Nohria e Tierney (1999) registrou que em algumas das organizações pesquisadas a estratégia estava centrada em um sistema de uso dos conhecimentos codificados e armazenados, típica do compartilhamento de conhecimentos explícitos.

A codificação tem por objetivo apresentar o conhecimento de forma a torná-lo acessível àqueles que precisam dele, transformando-o em código, não necessariamente de computador, para torná-lo compreensível, portátil e o mais organizado possível. A principal dificuldade nessa tarefa é codificar o conhecimento sem que haja perda de suas propriedades e sem transformá-lo em informação ou dados. É importante considerar que o conhecimento precisa de alguma estruturação, mas esta não pode estar em excesso (DAVENPORT; PRUSAK, 1998, p. 83-84).

O conhecimento estruturado e explícito de patentes, relatórios e documentos científicos não se torna utilizável simplesmente em decorrência de sua codificação. Ele precisa ser analisado e tornado acessível a pessoas que possam fazer algo com o mesmo e beneficiar a organização (DAVENPORT; PRUSAK, 1998, p. 103).

O mapeamento das fontes do conhecimento corporativo, que é tratado na seção 2.1.2.1, é uma parte importante do processo de codificação, além de ser nitidamente parte da infraestrutura de transferência do conhecimento (DAVENPORT; PRUSAK, 1998). Para Grotto (2003), o mapa do conhecimento organizacional é uma ferramenta formal que vem sendo utilizada nas organizações para promover o compartilhamento do conhecimento e facilitar seu acesso.

2.1.2.1 Mapeamento do conhecimento

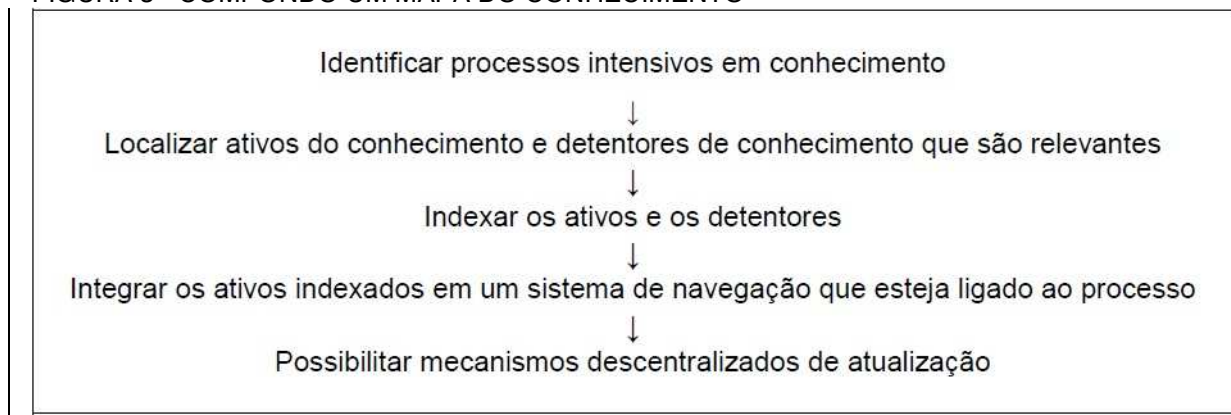
A principal função e o maior benefício do mapeamento do conhecimento (MC) “é mostrar às pessoas de dentro da organização para onde ir quando necessitarem do conhecimento” (DAVENPORT; PRUSAK, 1998). Ainda conforme os autores este mapa indica o conhecimento, mas não o contém, é um guia e não um repositório. O MC pode ser usado como ferramenta para avaliar o estoque de conhecimento corporativo, apresentando os pontos fortes a serem explorados e as lacunas a serem preenchidas.

As informações necessárias para a criação de um MC geralmente já existem nas organizações, embora de forma fragmentada e não documentada. O desenvolvimento do mapa envolve a localização de conhecimentos importantes dentro da organização e a sua publicação em algum tipo de lista ou quadro que mostre onde encontrá-los a fim de deixá-los acessíveis. Os mapas normalmente apontam para pessoas, documentos e bancos de dados, podendo, assim, remeter o pesquisador a pessoas ou a conhecimento estruturado ou a ambos (DAVENPORT; PRUSAK, 1998).

Para Schlesinger *et al.* (2008) o mapa do conhecimento está fortemente ligado ao processo de combinação, ou seja, sintetiza o conhecimento disperso na organização e disponibiliza a informação de onde encontrá-lo; o conceito de MC tem em sua essência os processos de socialização e externalização do conhecimento.

Segundo Probst, Raub e Romhardt (2002), os mapas do conhecimento aumentam a transparência e ajudam a identificar especialistas ou fontes de conhecimento; as informações que contêm podem ser informatizadas, organizadas e apresentadas visualmente com o auxílio de gráficos. A Figura 3 apresenta a proposta dos autores para os estágios em que os mapas do conhecimento devem ser produzidos.

FIGURA 3 - COMPONDO UM MAPA DO CONHECIMENTO

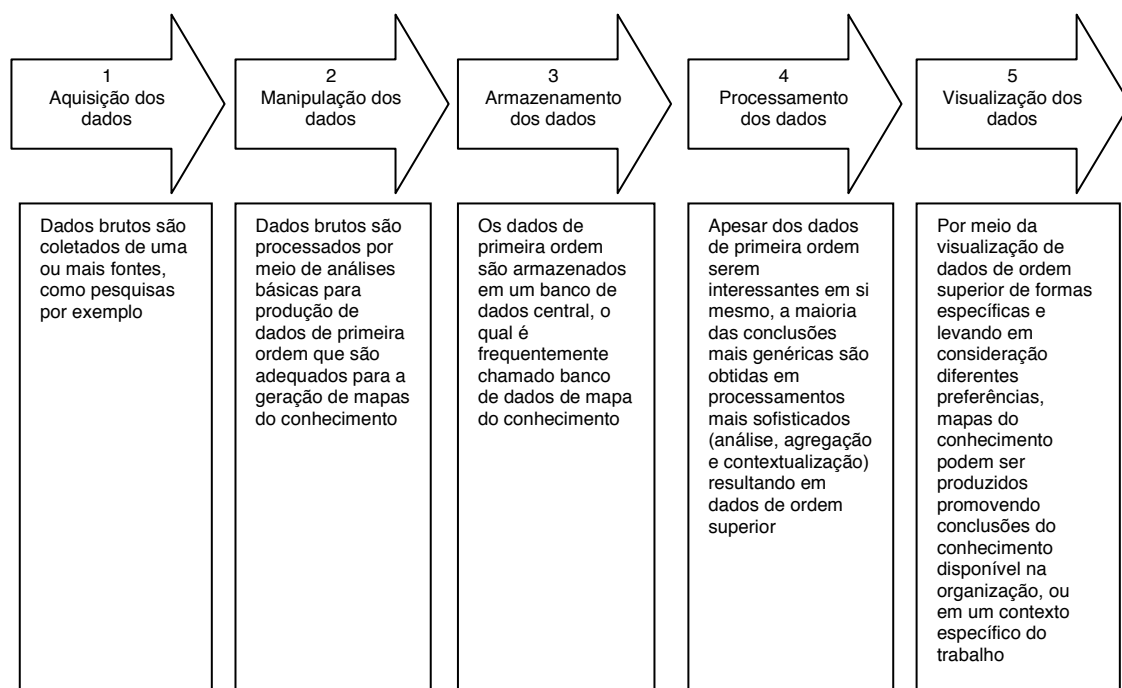


FONTE: Probst, Raub e Romhardt (2002, p.69)

Os autores propõem que a geração dos mapas se inicie com processos intensivos em conhecimento, que segundo Maldonado (2008) é um tipo particular de processo, essencial a GC “pois eles dependem fortemente do conhecimento embutido em atores, tarefas e atividades”. Já Davenport *et al.* (1996) identificam processos intensivos em conhecimento por meio da diversidade e incerteza de suas entradas e saídas, sendo o fluxo de eventos destes processos desconhecidos, uma vez que este pode evoluir durante a sua execução.

Ebener *et al.* (2006) apresenta os processos para desenvolvimento de um mapa do conhecimento, conforme Figura 4.

FIGURA 4 – PROCESSO DE MAPEAMENTO DO CONHECIMENTO



FONTE: Adaptado de Ebener *et al.* (2006) (adaptado de HUIJSEN; DRIESSEN; SWAAK (2004))

Na Figura 4 pode-se identificar as cinco etapas do processo de mapeamento do conhecimento:

- 1) aquisição dos dados: os dados brutos são obtidos de uma ou mais fontes de dados;
- 2) manipulação dos dados: os dados brutos são manipulados por meio de análises básicas para produzir dados de primeira ordem (dados de Primeira Forma Normal, este conceito é explicado na seção 2.3.2), que são adequados para geração de mapa do conhecimento;
- 3) armazenamento dos dados: os dados de primeira ordem são armazenados em uma base de dados centralizada;
- 4) processamento dos dados: apesar dos dados de primeira ordem serem adequados para geração de mapa do conhecimento, as conclusões mais elaboradas geralmente ocorrem com a aplicação de análises de agregação e contextualização sobre os dados de primeira ordem, gerando dados de ordem superior;
- 5) visualização dos dados: através da visualização dos dados de primeira ordem e ordem superior, considerando as preferências do usuário, podem ser elaborados mapas do conhecimento para dar indicações sobre o conhecimento que está na organização ou particularmente em um domínio de trabalho.

Segundo Grotto (2003), é importante considerar que o mapa facilita o acesso ao conhecimento, mas não garante que o compartilhamento seja efetivado. Esta ferramenta facilita a troca de conhecimentos tácitos ao mostrar quem o detém, mas é necessário contatar o detentor para que haja a transferência do conhecimento.

A tecnologia da informação pode ajudar a fazer com que o MC funcione; páginas amarelas ou um banco de dados de portadores do conhecimento podem ser colocados *online* para acesso na rede corporativa, permitindo que os usuários pesquisem por tópicos ou palavras-chave, o que facilita a localização e a comparação de fontes potenciais do conhecimento (DAVENPORT; PRUSAK, 1998).

Silva (2012) apresenta alguns exemplos de ações que podem ser tomadas nas empresas com o uso do mapa do conhecimento:

- a) estruturação de políticas de recursos humanos: os MC podem servir de instrumentos para identificar e definir conhecimentos necessários em perfis de candidatos, auxiliando no recrutamento. No desenvolvimento humano, pode ser utilizado para estabelecer as diretrizes de educação no processo de aprendizagem dos funcionários;
- b) identificação de colaboradores chave: ao auxiliar na identificação dos conhecimentos críticos relacionados ao negócio, o mapa permite visualizar as pessoas na organização que possuem conhecimentos que são imprescindíveis ao andamento de atividades consideradas críticas. Identificá-las com antecedência, possibilita a criação de ações que retenham estes talentos;
- c) estruturação de projetos de educação corporativa: ao optar pela implantação de um programa de educação, a empresa necessita identificar quais serão os conhecimentos a serem desenvolvidos em seus colaboradores, com o objetivo de alavancar as competências necessárias ao negócio;
- d) mapeamento de competências: para auxiliar no mapeamento do conjunto de conhecimentos, habilidades e atitudes, o MC apresenta-se como parte essencial do processo.

Um exemplo de representação do mapa do conhecimento é o software comercializado pela empresa CFEngine, com sede na Noruega e nos Estados Unidos. A forma de visualizar o mapa é por pesquisa por palavra-chave, conforme mostra a Figura 5, onde foi realizada a busca pela palavra: “webserver”.

A opção que está sendo apresentada na Figura 6 “Map” é uma exibição visual de uma pequena rede, onde constam os tópicos mais próximos relacionados. O tópico de pesquisa não pode ser o centro deste universo porque outros tópicos podem ser mais importantes. O diâmetro dos círculos é proporcional a quantidade de ligações ao tópico na árvore.

As outras opções que o software possui para visualização da informação são:

- “leads”: é uma explicação em forma textual para os vizinhos mais próximos ao tópico da busca;
- “references”: nesta opção o usuário visualiza *links* de documentos ou observações feitas sobre o tema da pesquisa;
- “same context”: esta opção retorna outros tópicos discutidos no mesmo contexto.

Outro exemplo do mapa de conhecimento pode ser visualizado na ferramenta proposta no trabalho de mestrado de Pirola (2002). Esta ferramenta permite gerar dez relatórios diferentes e realizar três tipos de buscas. Os relatórios podem ser gerados por: processos; pessoas-chave; capital estrutural e capital intelectual; capital de relacionamento e outros. Já as buscas podem ser realizadas por: pessoas-chave x atividade; pessoas-chave x habilidade e atividades x pessoas-chave. Um exemplo de relatório é apresentado na Figura 7.

FIGURA 7 – RELATÓRIO DE PESSOA-CHAVE

Home Relatórios Relatório de pessoas-chave					
Pessoa-Chave			Área da empresa		
Carlos Augusto			Finanças (Análise de Mercado)		
Atividades					
Análises Financeiras					
Relatório Quinzenal sobre o mercado financeiro					
Habilidades/Competências				Nível	
Finanças				Excelente	
Ferramentas Suite (Microsoft Office)				Bom	
Busca na Internet				Bom	
Redação				Bom	
Relacionamentos Internos (com outras pessoas da mesma empresa)					
Nome	Tipo	Grau de relacionamento		Imagem colaborador	Obs.
José Mauricio	Pessoal	Excelente		Muito bom	
Marco Antônio	Profissional	Péssimo		Péssimo	
Relacionamentos Externos (com pessoas de outras empresas)					
Nome	Empresa	Tipo	Grau de relacionamento	Imagem colaborador	Obs.
José Ricardo	TELEBRÁS	Profissional	Bom	Bom	
Maria Cândido	TELEMIG	Profissional	Indiferente	Bom	
Relacionamentos com outras empresas					
Empresa	Grau de relacionamento		Imagem colaborador		Obs.
TELEBRÁS	Bom		Bom		

FONTE: Pirola (2002)

No relatório mostrado na Figura 7 o gestor da área, pode identificar quais são os níveis de habilidade/competência da pessoa chave e seu grau de relacionamento com outra empresa, neste caso a empresa Telebrás. Estes dados podem ser importantes caso a empresa participe de um projeto em que alguma habilidade/competência específica seja necessária ou envolva uma iniciativa conjunta com a Telebrás.

2.1.2.2 Técnicas de mapa do conhecimento

Há algumas técnicas para mapear o conhecimento existente nas organizações. O Quadro 1 mostra um comparativo entre as técnicas.

QUADRO 1 - COMPARAÇÃO DAS TÉCNICAS DE MAPA DO CONHECIMENTO

	Páginas Amarelas	Análise de Fluxo de Informações	Análise de redes sociais	Mapeamento do processo do conhecimento	Mapeamento funcional do conhecimento
Ferramentas utilizadas para coleta de dados	Sistema de perguntas e respostas. Dicionário de conhecimentos e relatórios	Entrevistas e inventário de conhecimentos, extensivas pesquisas com diagramas de fluxo de informações	Questionário, Sociograma, Teoria dos grafos	Tempestade de ideias ou condução de entrevistas com os responsáveis pelos processos	Pesquisa e entrevistas
Ferramentas utilizadas para avaliação do mapa do conhecimento	Dicionário de conhecimento	Questionários, entrevistas e planilhas	InFlow, Krackplot e NetMiner	-	Observações, entrevistas e relatórios internos
Objetivos	Criar transparência na forma de localização do conhecimento por meio de registro das competências individuais em uma base de dados ou similar	Determinar quem está acessando qual recurso de informações e com que frequência	Descobrir padrões de interação entre os membros	Definir o conhecimento necessário, metas de decisão, o conhecimento disponível para apoiar os processos de negócio, rotas para acessar e recuperar o conhecimento e diferença entre o conhecimento atual e o necessário	Localizar áreas do conhecimento importantes. Identificar e caracterizar áreas de processos relacionados a "spots"
Abordagem para mapeamento do conhecimento	Baseado em projeto	Baseado em relacionamentos	Baseado em relacionamentos	Baseado em processos	Baseado em processos
Criação de mapas estatísticos ou dinâmicos	Estatístico	Estatístico	Dinâmico	Dinâmico	Dinâmico
Suporte tácito ou explícito	Explícito	Tácito	Tácito	Explícito, Tácito	Explícito, Tácito

FONTE: Adaptado de Jafari *et al.* (2009)

Tendo como base Jafari *et al.* (2009), são apresentadas as cinco técnicas para mapeamento do conhecimento que constam no Quadro 1:

- 1) páginas amarelas: é um conjunto de dados e informações sobre as pessoas de uma organização, é uma maneira de identificar quem sabe o que na organização. Seu objetivo é facilitar a comunicação e o compartilhamento de conhecimento entre os indivíduos e grupos de pessoas nas organizações;
- 2) análise de fluxo de informações: identifica quem está acessando qual tipo de informação e quantas vezes, por meio de uma pesquisa no processo funcional da organização e redes informais. Este processo pode ser posto em prática pelo uso de programas de computadores e por questionários e entrevistas com as pessoas;
- 3) análise de redes sociais: pode ser definida como um mapa de relações e fluxos entre pessoas, grupos, organizações, computadores ou outras informações. Os nós (ou nodos, ou ainda vértices) de cada rede são as pessoas e grupos e as ligações mostram as relações ou fluxos entre os nós. Esta técnica busca entender como o conhecimento tácito foi transmitido e como este compartilhamento pode ser melhorado;
- 4) mapeamento do processo do conhecimento: é um método de análise para definir o conhecimento necessário e o conhecimento disponível para auxiliar o processo de negócio da organização, onde também este processo é analisado a fim de identificar marcos de decisão, requisitos de conhecimento, formas de acesso e recuperação do conhecimento por meio de pessoas, tecnologia e lacunas entre as competências e habilidades;
- 5) mapeamento funcional do conhecimento: é semelhante ao mapeamento anterior, mas este suporta mais detalhes, pois esta técnica permite considerar o conhecimento individual e contatos sociais que estão relacionados à posição da pessoa na organização. Este mapa cria um diretório organizacional de habilidades, conhecimentos, relações individuais e recursos.

A seção 2.2 trata sobre a visualização. Conceituando a técnica utilizada para representar o mapa do conhecimento a fim de permitir que o usuário realize as análises de forma simples.

2.2 VISUALIZAÇÃO

A visualização é o processo que apresenta a informação/conhecimento de forma visual, e tem sido usada para analisar e mostrar grandes volumes de dados multidimensionais. Ela permite, diferentemente dos métodos estatísticos, visualizar os resultados sem necessariamente saber que tipo de fenômeno deve ser analisado. A visualização é mais do que um método computacional. É um processo que permite aos usuários observar a informação, possibilitando uma análise que não seria possível sem o auxílio de componentes visuais (GERSHON, 1994). A visualização é baseada em técnicas que transformam os dados em imagens, podendo ser em um espaço unidimensional (1D), bidimensional (2D), tridimensional (3D) ou mesmo multidimensional, n -dimensional (nD), caso a visualização de informações contenha múltiplas variáveis a serem codificadas em uma única estrutura visual (CARD; MACKINLAY; SHNEIDERMAN, 1999).

De acordo com a natureza dos dados que são manipulados, utilizam-se termos mais específicos como visualização científica e visualização da informação. Em ambas as áreas, são criados modelos gráficos ou representações visuais dos dados.

Enquanto a visualização científica se preocupa em analisar dados de natureza física, que geralmente possuem uma característica espacial, a visualizada da informação está preocupada em estudar dados abstratos que não possuem referências espaciais e cuja complexidade é aumentada devido às grandes quantidades de informações. Para Rhyne (2003), as fronteiras destas áreas não são nítidas e nem está claro que haja vantagens nesta separação.

Já a diferença entre a visualização da informação e visualização do conhecimento (VC) está na maneira de usar as capacidades visuais; enquanto a visualização da informação visa explorar dados abstratos e criar novos *insights*, a

VC, além disso, tem por objetivo melhorar a transferência de conhecimento entre as pessoas (SASIETA, 2011).

2.2.1 Representação visual

As representações visuais correspondem a modelos gráficos, figuras ou imagens utilizadas para mapear graficamente conjuntos de dados a serem analisados e/ou explorados. Conforme Spence (2001), o grande desafio da visualização é o de criar metáforas visuais que efetivamente representem as informações e possibilitem maneiras satisfatórias de explorar e manipular os dados representados.

Conforme é apresentado no Quadro 2, as representações visuais podem ser classificadas em uma das seguintes categorias: gráficos; ícones, glifos e objetos geométricos; mapas e diagramas (FREITAS *et al.*, 2001).

QUADRO 2 - CLASSES DE REPRESENTAÇÕES VISUAIS

Classe	Tipo	Utilização
Gráficos 2D, 3D	Pontos	Representação da distribuição dos elementos no espaço domínio, representação da dependência/correlação entre atributos.
	Circulares	
	Linhas	
	Barras	
	Superfícies (para 3D)	
Ícones	Elementos geométricos 2D ou 3D	Representação de entidades num contexto, representação de grupos de atributos de diversos tipos.
Glifos	diversos	
Objetos geométricos		
Mapas	de pseudo-cores	Representação de campos escalares ou de categorias.
	de linhas	Representação de linhas de contorno de regiões, isovalores.
	de superfícies	Idem, no espaço 3D.
	de ícones, símbolos diversos	Representação de grupos de atributos (categóricos, escalares, vetoriais, tensoriais).
Diagramas	Nodos e arestas	Representação de relacionamentos diversos: É-um, É-parte-de, Comunicação, Sequência, Referência, etc.

FONTE: Freitas *et al.* (2001)

O Quadro 2 apresenta, além dos gráficos tradicionalmente utilizados para apresentação de resultados (pontos, linhas, barras, torta e de frequência), outros que possibilitam observar as relações entre atributos, uma série de representações gráficas utilizadas para codificar, por meio de elementos visuais (cores ou símbolos geométricos) tanto valores como relacionamentos entre entidades e/ou elementos de dados.

2.2.2 Categorização das técnicas de visualização

Existem diversas técnicas de categorizações propostas, entre elas: as taxonomias apresentadas por North (1998) e Chi (2000), as classificações adotadas em Card, Mackinlay e Shneiderman (1999) e Olive (1999) e as categorizações para técnicas de visualização de dados multidimensionais sugeridas por Keim (1997).

Optou-se por descrever a categorização adotada por Card, Mackinlay e Shneiderman (1999), por identificar que esta apresenta uma discriminação adequada dos tipos básicos de estruturas visuais para esta pesquisa, agrupando as aplicações em quatro níveis.

No primeiro nível constam as ferramentas de visualização que proveem aos usuários acesso visual a coleções de informações externas a seu meio imediato, como dados disponíveis na Internet ou em bases de dados *on-line*. No segundo nível constam as ferramentas que visam apoiar os usuários na execução de tarefas, criando representações visuais altamente intuitivas para cumprir este objetivo. O terceiro nível contém as informações visuais de extração do conhecimento, que descrevem representações visuais de determinados dados e um conjunto de controles para interagir com essas representações, de maneira que os usuários possam determinar e extrair relacionamentos dos dados. Já o quarto nível, abrange as aplicações que apresentam objetos visualmente realçados, com o propósito de revelar mais informações sobre um objeto usando uma forma visual intrínseca.

Retomando o terceiro nível, as ferramentas visuais do conhecimento são adicionalmente categorizadas conforme o tipo de estrutura visual adotada, a qual se refere à dimensão das representações de dados utilizada. Entre as estruturas básicas estão:

- a) físicas, ou naturais: possuem uma correspondência direta com objetos, fenômenos ou posições, sendo relacionadas à representação de dados típicos da visualização científica. Esta categoria abrange todas as técnicas destinadas à construção e visualização de representações 3D de objetos do mundo físico, como: moléculas e proteínas, imagens do corpo humano e mapas topográficos;
- b) 1D, 2D, 3D: dizem respeito a representações visuais que codificam informações por meio do posicionamento de símbolos ou marcas em espaço unidimensional, bidimensional ou tridimensional. As estruturas unidimensionais podem ser observadas na representação de dados temporais, como em linhas de tempo. Já as estruturas visuais de 2D e 3D são normalmente identificadas em gráficos tradicionais (de barras, colunas, linhas, círculos) e utilizadas para representar dados físicos e, também, informações abstratas;
- c) multidimensionais: caracterizam-se por ambientes de visualização que manipulam dados, normalmente abstratos, contendo vários atributos a serem mapeados em estruturas visuais de 1D, 2D ou 3D. Enquadram-se nesta categoria técnicas de projeção geométrica, iconográficas e orientadas a *pixel*;
- d) árvores e redes: abrangem todas as técnicas que estabelecem uma hierarquia para a exibição de dados, independentemente de estes apresentarem ou não uma estrutura hierárquica inerente. Consistem, comumente, de nodos representando dados e ligações indicando o relacionamento entre os nodos. Estas estruturas são empregadas para representar, por exemplo, taxonomias, organizações, gerenciamento de tráfego ou estrutura de informação na Internet, que são mais bem descritas e exemplificadas na seção 2.2.3.

2.2.3 Técnicas de visualização hierárquicas

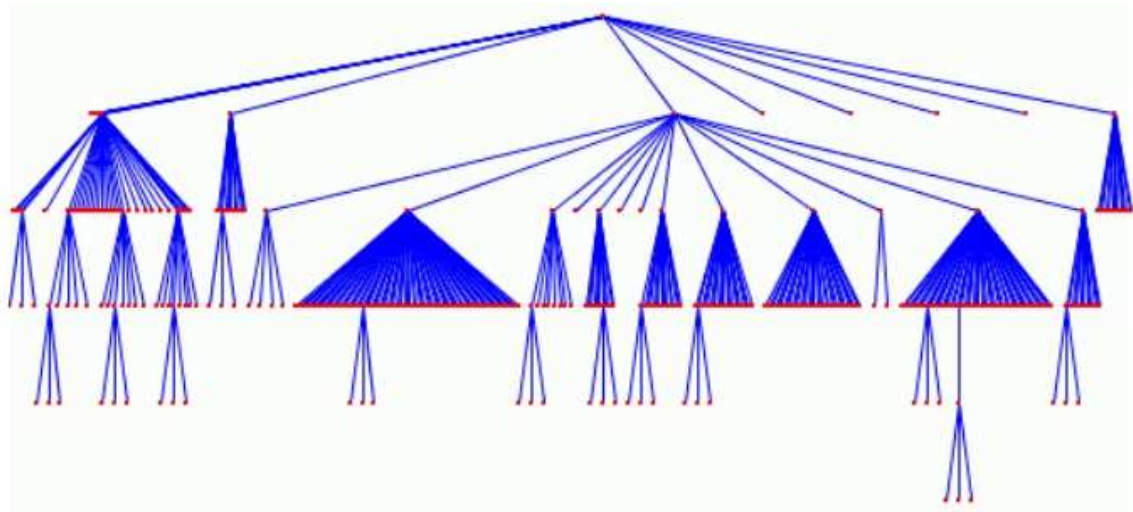
Geralmente, as técnicas hierárquicas são aplicadas em dados cuja própria natureza apresenta uma correlação explícita entre níveis e subníveis, por exemplo,

as pastas ou diretórios em um sistema de arquivos. Desta forma, o espaço n -dimensional dos dados é dividido em subespaços que são organizados uns dentro dos outros, sendo exibidos de forma hierárquica.

Herman, Melançon e Marshall (2000) realizaram uma revisão dos métodos para representação de grafos e árvores e destacam que o tamanho do grafo é uma questão chave na visualização, tendo em vista que grandes grafos, com muitos elementos, podem comprometer sua visibilidade e usabilidade por tornar inviável a distinção entre nós e arestas.

A Figura 8 apresenta a disposição da árvore clássica.

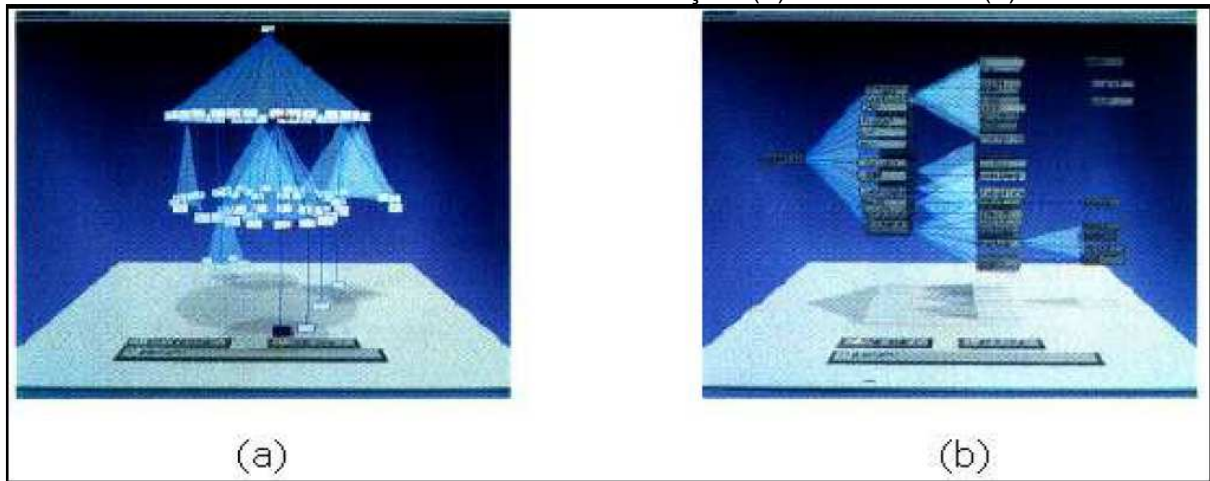
FIGURA 8 - LEIAUTE DE ÁRVORE MODERADAMENTE GRANDE



FONTE: Herman, Melançon e Marshall (2000)

A Figura 8 apresenta a estrutura básica posicionando os nós filhos abaixo de seu ancestral comum. O algoritmo de Reingold e Tilford (1981) é o mais adotado esquema da técnica de disposição em árvore e pode ser adaptado para ser apresentado de cima para baixo, bem como da esquerda para a direita, como as técnicas apresentadas a seguir.

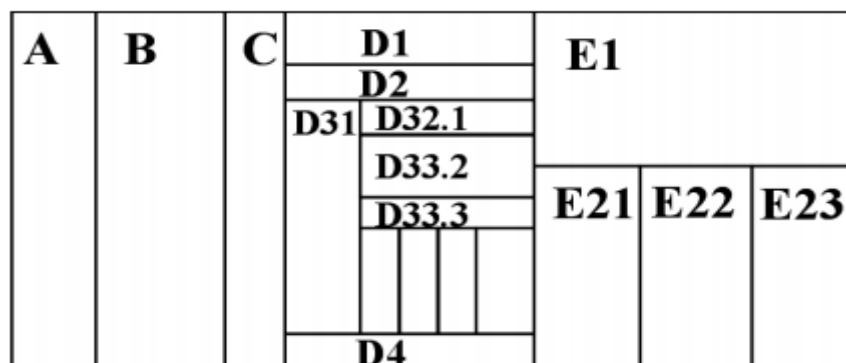
Técnicas como: *Cone Trees* e *Cam Trees* são a representação tridimensional das árvores em 2D. Estas técnicas se diferenciam apenas pela posição da árvore de cones formada; enquanto a *Cone Trees* é representada na horizontal, a *Cam Trees* é representada na vertical (ROBERTSON; MACKINLAY; CARD, 1991), conforme ilustra a Figura 9.

FIGURA 9 - TÉCNICAS HIERARQUICAS DE VISUALIZAÇÃO (A) *CONE TREE* E (B) *CAM TREE*

FONTE: Robertson, Mackinlay e Card (1991)

A construção da árvore na técnica proposta na Figura 9 é realizada a partir de um nó raiz localizado no vértice de um cone. Todos seus filhos, ou folhas da árvore são igualmente espaçados e posicionados na base deste cone. Este processo se repete para cada nó da árvore que possui filhos. Em cada nível, a altura e o diâmetro da base dos cones são então recalculados para que toda a informação esteja visível (FREITAS *et al*, 2001).

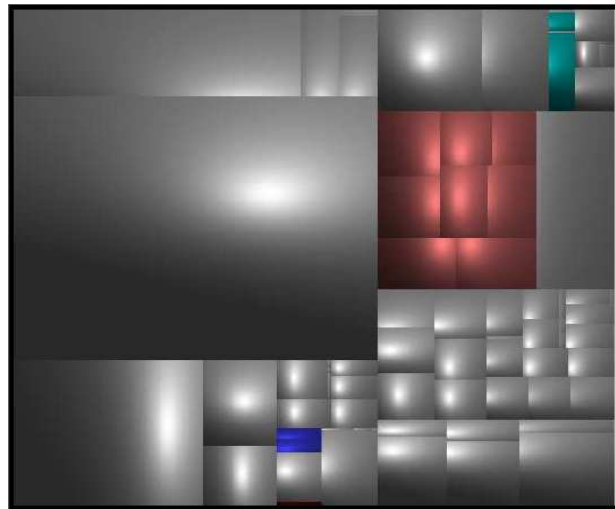
Outra abordagem é a técnica *Treemap*, adotada por Johnson e Schneiderman (1991), que utiliza o espaço de tela para representar elementos de informação, ao invés de usar objetos geométricos. Esta técnica mapeia as informações dividindo a tela em partes para representar os diretórios e subdivide estas partes para representar os subdiretórios e assim por diante. Uma desvantagem deste método é a dificuldade de identificar visualmente os diferentes níveis do *Treemap* quando a hierarquia se torna muito profunda. A Figura 10 representa o uso desta técnica em um esquema básico.

FIGURA 10 – ESQUEMA BÁSICO DA TÉCNICA *TREEMAPS*

FONTE: Johnson e Shneiderman (1991)

Buscando aumentar a satisfação do usuário com a aparência dos Treemaps e tornar a informação da hierarquia mais intuitiva para identificar os níveis, Wijk e Wetering (1999) desenvolveram uma técnica denominada Cushion Treemaps. Na Figura 11, pode ser observado que o algoritmo desta técnica usa a iluminação dos pixels e uma função de altura para representar os diversos níveis da hierarquia. Esta função de altura é acionada para destacar retângulos menores.

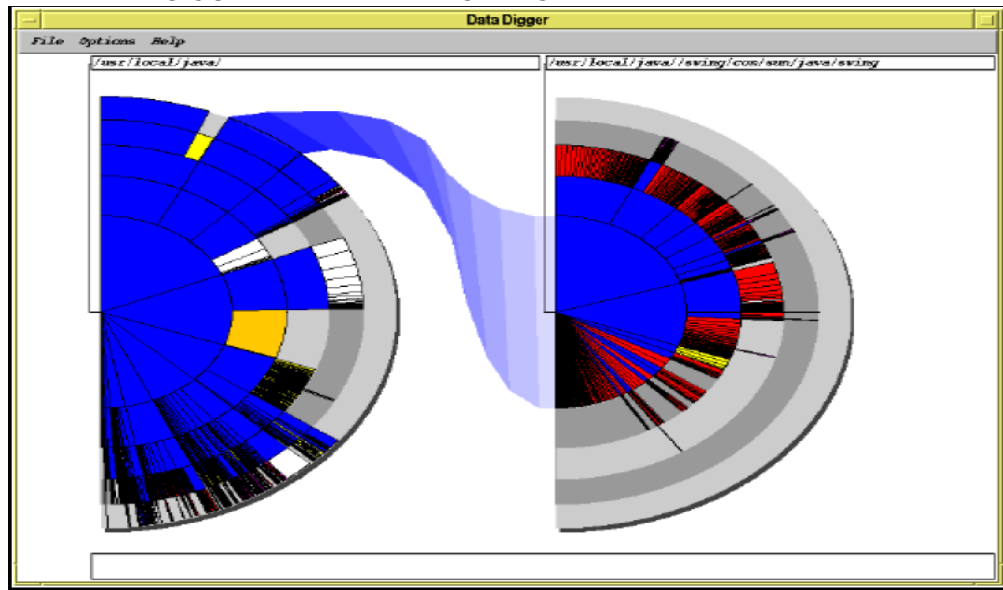
FIGURA 11 - TÉCNICA *CUSHION TREEMAPS*



FONTE: Wijk e Wetering (1999)

A técnica *Information Slices* usa discos semicirculares para representar hierarquias, conforme ilustra a Figura 12. Esta técnica permite que dados sejam vistos de forma compacta com vários níveis em duas dimensões (ANDREWS; HEIDEGGER, 1998). Cada disco representa uma hierarquia de múltiplos níveis. Em cada nível da hierarquia, os filhos são dispostos de acordo com os valores dos dados. Para hierarquias maiores uma série de discos em cascatas pode ser utilizada. Um segundo semicírculo pode ser usado para representar níveis com mais detalhes.

FIGURA 12 - TÉCNICA *INFORMATION SLICES* COM SEMICÍRCULO AUXILIAR PARA APRESENTAR NÍVEIS COM DETALHAMENTO MAIOR



FONTE: Andrews e Heidegger (1998)

No que se refere a grafos, a ideia é visualizar dados usando técnicas que mapeiem as características, como direcionado ou não-direcionado, de um dado grafo, de maneira clara e rápida. Os métodos podem ser subdivididos segundo a dimensionalidade visual da representação: 2D ou 3D.

Conforme Menezes (2012), todo conjunto de dados que possui qualquer tipo de relação pode ser representado por um grafo. Os nós do grafo representam os elementos e as arestas representam as relações entre esses elementos. A visualização em grafos pode ser aplicada a diversas áreas: busca de arquivos em árvores, mapeamento do histórico em sites da Internet, redes sociais e outros.

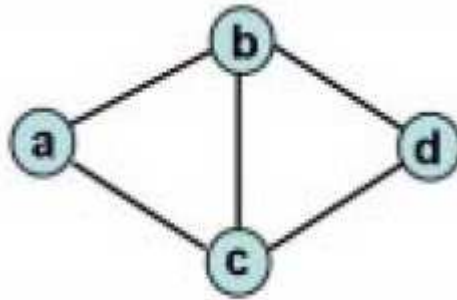
2.2.4 Conceito de grafos

Um grafo G consiste em um conjunto de vértices $V(G)$ e um conjunto de arestas $A(G)$, onde cada aresta é um par não ordenado de vértices (CHUNG ;LU, 2006).

Por exemplo, a Figura 13 apresenta um grafo $G = (V(G), A(G))$ ou $G(V, A)$ definido como a seguir:

$$V(G) = \{a; b; c; d\};$$

$$A(G) = \{\{a; b\}; \{a; c\}; \{b; c\}; \{b; d\}; \{c; d\}\}.$$

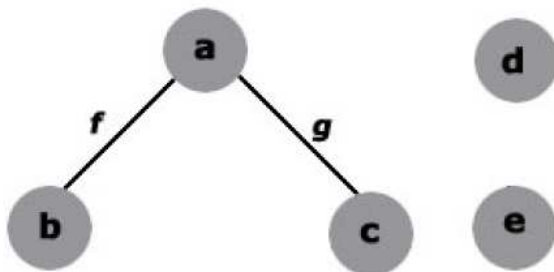
FIGURA 13 – UM GRAFO $G(V,A)$ 

FONTE: Chung e Lu (2006)

Segundo Saito (2003) grafos podem ser utilizados para modelar sistemas e estruturas onde há um relacionamento entre os objetos e estes são importantes para obtenção do resultado final. Normalmente na representação gráfica as vértices, ou nós são representados por círculos ou caixas e são conectados pelas linhas que indicam a existência de relacionamentos entre os objetos (arestas).

Em um grafo os elementos de V são representados por nomes, letras ou números. Já as arestas são desenhadas apenas quando existe a ligação entre os elementos, caso a ligação não exista a linha não é representada. Um vértice que não possua nenhuma aresta incidente é chamado de incidente, enquanto que um par de arestas incidente sobre um vértice comum é chamado de adjacente. A Figura 14 apresenta um exemplo onde as arestas “d” e “e” são vértices isolados e o conjunto de arestas “f” e “g” é adjacente, tendo em vista que, “f” e “g” incidem sobre a aresta/nó “a” (SAITO, 2003).

FIGURA 14 – EXEMPLO DE UM GRAFO COM VÉRTICES ISOLADOS



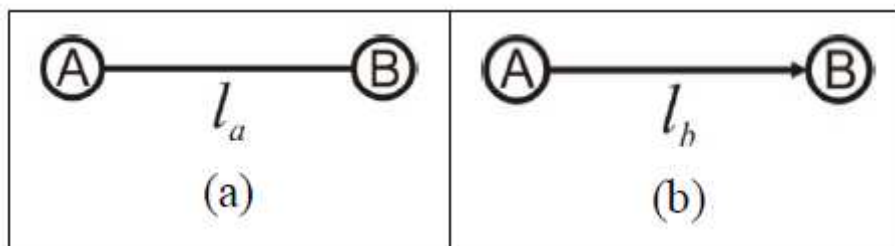
FONTE: Saito (2003)

Outro conceito abordado por Menezes (2012) é que um nó é incidente em uma aresta se ele for um dos nós definidos ou como origem ou como destino da linha, como apresenta na Figura 15.

Um grafo pode indicar diferentes tipos de relacionamentos entre os objetos, neste caso chamado multi-relacional, sendo que, esses relacionamentos podem ser representados por meio de desenhos diferentes. Por exemplo, em uma rede social científica podem ser desenhadas linhas contínuas para representar os relacionamentos de co-autoria entre os pesquisadores e usar linhas pontilhadas para identificar os relacionamentos de participação no desenvolvimento do mesmo projeto (MENEZES, 2012).

Ainda conforme Menezes (2012) as ligações em um grafo podem ser do tipo não direcionadas ou direcionadas, como representado nas relações l_a e l_b , da Figura 15. Nas relações não direcionais os dois atores cooperam com aquele tipo de relação. Por outro lado, nos grafos direcionados apenas o elemento oposto à seta é o que exerce a relação. Desta forma, caso os dois elementos tenham a relação é necessário representar com duas linhas com setas em sentidos opostos ou uma linha com duas setas.

FIGURA 15 – TIPOS DE ARESTAS NO GRAFO



FONTE: Menezes (2012)

As técnicas de visualização apresentadas podem ser usadas para auxiliar, ou serem auxiliadas pela descoberta de conhecimento em base de dados, que será tratada na seção 2.4, mas primeiramente são apresentados conceitos básicos referentes a banco de dados na seção 2.3.

2.3 BANCO DE DADOS

Segundo Guimarães (2008), um banco ou base de dados – BD, trata-se de uma coleção de dados ou informações relacionadas entre si, estas representam aspectos do mundo real e o que desejamos armazenar para uso futuro. Yong (1983,

p.18) descreve como “uma base de dados constituída de um conjunto de arquivos relacionados entre si”.

Alguns dos conceitos essenciais relacionados a banco de dados, de acordo com a abordagem de Yong (1983), estão listados a seguir:

- tabela: é um conjunto de dados organizados de forma bidimensional com linhas e colunas, onde as linhas representam os registros e as colunas representam campos;
- entidade: utilizando o conceito de entidade do contexto da modelagem entidade-relacionamento (MER), esta é vista como um modelo computacional que representa uma pessoa, um objeto ou algo real que precisa ser representado ou armazenado e processado nos sistemas computacionais (SILBERSCHATZ; KORTH; SUDARSHAN, 2006);
- relacionamento: a associação que existe entre as informações de duas ou mais tabelas é chamada de relacionamento. Sendo que, a ligação entre as tabelas pressupõe a existência de um atributo comum;
- índice: os índices são utilizados para manter a ordem dos registros de uma tabela segundo um ou mais de seus atributos. Os atributos que fazem parte de um índice são chamados de chave;
- chave: é um atributo de uma tabela que faz parte de um índice. Quando uma chave é formada por dois ou mais atributos, esta é denominada como chave composta. O identificador único que é um elemento que identifica e preserva a identidade e unicidade de um registro, é chamado de id ou código e é utilizado para ser a chave primária (*primary key* – PK) dessa tabela. O id de uma tabela pode ser usado em uma segunda tabela para fins de relacionamento, nesta situação o atributo da segunda tabela é chamado de chave estrangeira (*foreign key* - FK);
- cardinalidade: a cardinalidade é utilizada para descrever o grau de relacionamento entre duas tabelas, geralmente o relacionamento é : 1x1, 1xN ou NxN. A descrição deste relacionamento está na seção 2.3.3.

Os bancos de dados modernos geralmente atendem a um conjunto de requisitos conhecidos pelo mnemônico em inglês ACID, originado das palavras atomicidade, consistência, isolamento e durabilidade.

Atomicidade significa que cada transação (conjunto de instruções encadeadas para serem executadas no BD) deve executar com sucesso todas as instruções, caso contrário todas as instruções devem ser desfeitas.

Consistência significa que cada transação deve utilizar dados originários de um estado válido do banco de dados. A validade dos dados está associada ao atendimento de regras disponíveis no BD como unicidade ou não do dado, permissão ou não de valores nulos, validade de relacionamentos e a manutenção do estado válido.

Isolamento significa que no caso de 2 ou mais transações ocorrerem simultaneamente, cada qual deve executar de forma independente das outras.

Durabilidade significa que a fim de uma transação executada com sucesso, o resultado das instruções da transação estão disponíveis permanentemente no banco, ou seja, não deve ocorrer perda destes resultados.

Para atender a definição e requisitos apresentados é necessário a construção de programas complexos e elaborados. Esta coleção de programas é chamada de Sistema Gerenciador de Banco de Dados (SGBD).

Um SGBD deve proporcionar técnicas e ferramentas para definição e descrição dos dados armazenados, preferencialmente de forma a ocultar os detalhes técnicos envolvidos. A atividade de definição e descrição dos dados chama-se modelagem de dados.

2.3.1 Modelos de dados

A modelagem de dados de um BD envolve a análise dos requisitos dos usuários do BD e a forma como estes usuários enxergam externamente os dados. (GUIMARÃES, 2008).

A estrutura de banco de dados é apoiada pelo modelo de dados, que é definido como uma coleção de ferramentas conceituais com o objetivo de descrever dados e suas relações, semântica de dados e restrições de consistência (SILBERSCHATZ; KORTH; SUDARSHAN, 2006).

Para os autores, os modelos de dados podem ser agrupados em quatro categorias:

- 1) modelo relacional: este modelo utiliza uma coleção de tabelas para representar os dados e as relações entre eles. Cada tabela possui diversas colunas e cada coluna possui um nome único;
- 2) modelo entidade-relacionamento (MER): este modelo está baseado na percepção de um mundo real que consiste em uma coleção de objetos, chamados de entidades e as relações entre eles;
- 3) modelo de dados baseado em objeto: pode ser visto como uma extensão do MER com noções de encapsulamento, métodos e identidade de objeto;
- 4) modelo de dados semiestruturado: permite que os dados sejam especificados para que itens de dados individuais do mesmo tipo possam ter diferentes conjuntos de atributos. Sendo o oposto dos modelos de dados citados anteriormente, nos quais todos os itens de dados de um determinado tipo precisam ter o mesmo conjunto de atributos (SILBERSCHATZ; KORTH; SUDARSHAN, 2006, p. 5).

Para Guimarães (2008) de acordo com o nível de detalhamento necessário na modelagem de dados pode gerar diferentes tipos de modelo, como o modelo conceitual, lógico e físico.

O modelo conceitual é uma representação mais abstrata dos requisitos obtidos junto aos usuários e é totalmente independente do SGBD utilizado. Este modelo oferece a descrição dos diversos tipos de dados requeridos, os relacionamentos existentes entre os dados e algumas regras de consistência.

O modelo lógico transmite os conceitos disponíveis no modelo conceitual em um formato mais adequado ao SGBD. Este modelo se mantém na fronteira entre um modelo independente do SGBD e um modelo específico para o SGBD.

O modelo físico corresponde a organização interna do SGBD e apresenta características específicas do SGBD selecionado.

Peter Pin-Shan Chen no ano 1976 apresentou uma proposta de modelagem conceitual conhecida por modelo entidade-relacionamento (MER). Esta proposta foi rapidamente adotada pelos usuários devido a facilidade de compreensão e a capacidade de formalizar aspectos do BD. A modelagem conceitual utilizando MER consiste em elaborar diagramas entidade-relacionamento que descrevem os dados do BD e os seus relacionamentos (GUIMARÃES, 2008).

2.3.2 Modelo entidade-relacionamento

No âmbito do modelo entidade-relacionamento, uma entidade é um objeto ou ente do mundo real que existe e cujas características deseja-se armazenar. Segundo Guimarães (2008), esta entidade pode ser física (uma pessoa, um carro) ou abstrata (um departamento, um projeto, um idioma).

As entidades possuem características que deseja-se armazenar. Estas características são conhecidas por atributos. Um atributo de uma entidade carro pode ser, por exemplo, a cor da pintura.

Quando deseja-se armazenar diversas entidades que compartilham as mesmas características, agrupa-se as entidades em um conjunto de entidades (CE). Um CE é constituído por zero, uma ou mais entidades distintas entre si.

Em geral, é necessário identificar de forma única as entidades dentro de um CE. Para isto define-se um atributo chave da entidade de forma que é possível identificar a entidade por meio deste atributo-chave.

2.3.3 Relacionamentos

Um relacionamento entre dois CEs é uma lista, ou um conjunto de pares de entidades. Cada par representa uma associação entre uma entidade de um CE com outra entidade de outro CE e que estabelece uma interdependência entre essas entidades (GUIMARÃES, 2008).

Os relacionamentos podem ser classificados de acordo a cardinalidade descrita no Quadro 3.

QUADRO 3 – DESCRIÇÃO DOS RELACIONAMENTOS PELA CARDINALIDADE

Cardinalidade	Descrição
N : 1 ou Nx1	Cenário em que uma entidade de um CE pode estar associado a várias entidades de um outro CE. Exemplo: uma entidade departamento do CE departamentos pode estar associada a diversas entidades funcionários de um CE funcionários.
1 : 1 ou 1x1	Cenário em que uma entidade de um CE só pode estar associada a uma única entidade de outro CE. Exemplo: departamentos são gerenciados por funcionários de forma que dado um departamento existe um único funcionário que o gerencia e dado um funcionário, ele pode ser gerente de um único departamento.

Cardinalidade	Descrição
N : N ou NxN	Cenário em que uma entidade de um CE pode estar associada a várias entidades de outro CE e vice-versa. Exemplo: um funcionário pode participar de diversos projetos e um projeto pode ter a participação de diversos funcionários.

FONTE: Adaptado de Guimarães (2008)

2.3.4 Normalização de dados relacionais

Uma importante etapa na elaboração da modelagem relacional é o processo de normalização, que tem por objetivo reduzir a redundância dos dados. Para Yong (1983, p. 125) “o processo de transformação de relações não normalizadas às formas mais ótimas de expressão é conhecido como normalização”.

Existem cinco níveis de formas normais, representadas pelos símbolos: 1FN, 2FN, 3FN, 4FN e 5FN. As três primeiras são as mais conhecidas e aplicadas. É importante salientar que cada nível depende do nível anterior e é uma especialização do caso anterior, portanto a Segunda Forma Normal (2FN) é uma especialização da 1FN e obedece às relações da 1FN.

Uma relação R está na 1FN se todos os domínios considerados contenham somente valores atômicos, ou seja, valores que não têm relações (YONG, 1983, p.127).

Um exemplo para esta definição apresentada por Yong (1983) é que uma tabela R está na Primeira Forma Normal (1FN) se todos os campos, ou atributos considerados contenham somente valores atômicos, ou seja, o conteúdo previsto para o campo não pode ser mais dividido em dois ou mais campos. Desta forma, uma tabela que armazene dados de pesquisadores, por exemplo, poderá ser denominada de pesquisador e pode ter atributos como: matrícula, nome e endereço. A princípio, os campos matrícula e nome estão na forma atômica, isto é, não podem ou não precisam ser divididos em mais campos. O campo “endereço”, contudo, pode ser dividido em logradouro, número, complemento, bairro, CEP, cidade e UF. Afirma-se, portanto, que o campo é composto ou que é uma relação ou que não se encontra na forma atômica.

Ainda segundo Yong (1983) uma relação R está na 2FN se os atributos não chave forem totalmente dependentes dos atributos-chave, ou seja, dependentes da

chave primária. E, está na Terceira Forma Normal (3FN) se “os atributos dependentes, não são dependentes transitivos do atributo-chave”.

2.4 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Segundo Fayyad *et al.* (1996, p. 3-4) o termo descoberta do conhecimento em base de dados (*Knowledge Discovery in Databases* – KDD) foi cunhado em 1989 para se referir ao amplo processo de encontrar o conhecimento em dados e para enfatizar a aplicação em "alto nível" de determinados métodos de mineração de dados. O termo mineração de dados - MD (*data mining*) é mais utilizado por estatísticos, analistas de dados e pela comunidade de sistemas de informações gerenciais, enquanto KDD tem sido mais utilizado pelos pesquisadores de inteligência artificial e de aprendizagem de máquina.

Ainda baseado nestes autores, esta pesquisa irá adotar que o termo KDD se refere ao processo global de descoberta de conhecimento a partir de dados enquanto mineração de dados se refere à aplicação de algoritmos para a extração de padrões de dados sem os passos adicionais do processo de KDD.

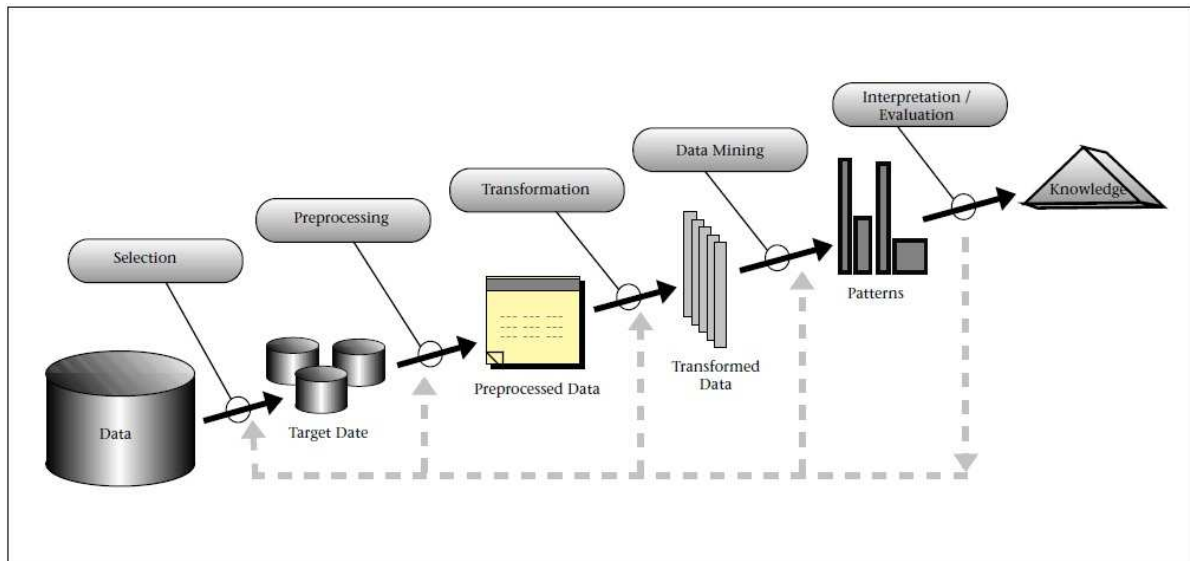
Conforme Frawley, Piatetsky-Shapiro e Matheus (1991, p. 3-5) descoberta de conhecimento é a extração não trivial de informação implícita, previamente desconhecida e potencialmente útil a partir de dados, possuindo quatro características principais para o KDD:

- a) linguagem de alto nível: o conhecimento descoberto é representado em uma linguagem de alto nível, que não precisa ser usada diretamente por pessoas, mas sua expressão precisa ser compreensível por usuários humanos;
- b) precisão: as descobertas devem retratar fielmente o conteúdo do banco de dados;
- c) resultados interessantes: o conhecimento descoberto é interessante de acordo com as tendências definidas pelo usuário, considerando que ser interessante implica que os padrões são novos e potencialmente úteis;
- d) eficiência: o processo de descoberta é eficiente, tempos de execução de bases de grande porte são previsíveis e aceitáveis.

2.4.1 Processo de KDD

O processo de KDD é interativo e iterativo, envolvendo numerosos passos. A Figura 16 apresenta uma visão geral dos passos que são compreendidos pelo processo.

FIGURA 16 - VISÃO GERAL DOS PASSOS QUE COMPOEM O PROCESSO DE KDD.



FONTE: Fayyad et al. (1996 b)

A seguir serão detalhadas as etapas do processo de KDD apresentado por Fayyad et al. (1996, p. 10-11):

- 1) desenvolver um entendimento sobre o domínio da aplicação, identificar o conhecimento relevante existente e os objetivos do usuário final;
- 2) criação de um conjunto de dados alvo: seleção do conjunto de dados ou foco em um subconjunto de variáveis ou dados, nos quais a descoberta será aplicada;
- 3) limpeza dos dados e pré-processamento: operações básicas como remoção de ruídos ou discrepâncias, se preciso, coleta das informações necessárias para tratamento dos ruídos, decisão sobre quais estratégias serão usadas para tratar dados ausentes, manipulação de informações temporais e alterações conhecidas;
- 4) redução e projeção dos dados: encontrar características úteis para representar depende do objetivo da tarefa; utiliza métodos de redução dimensional ou transformação para reduzir o número efetivo de variáveis consideradas ou para encontrar representações invariantes dos dados;

- 5) escolha da tarefa de mineração de dados: decidindo se a meta do processo de KDD é classificação, regressão, agrupamento, etc;
- 6) escolha do(s) algoritmo(s) de mineração de dados: selecionando método(s) para serem utilizados na procura por padrões nos dados. Isto inclui decidir quais modelos ou parâmetros podem ser apropriados e ajustar um método de MD em particular com os critérios gerais do processo KDD;
- 7) mineração de dados: procurar por padrões de interesse em uma forma particular de representação ou um conjunto de representações: regras de classificação, árvores de decisão, regressão, agrupamento e assim por diante;
- 8) interpretação dos padrões minerados e possível retorno para qualquer dos passos 1 a 7 para nova iteração;
- 9) consolidação do conhecimento descoberto: incorporação deste conhecimento no sistema de performance, ou simplesmente documentando e reportando para as partes interessadas. Isto também inclui checar e resolver conflitos potenciais com o conhecimento adquirido ou extraído anteriormente.

O passo 7, referente a mineração de dados, é a principal etapa no processo de KDD, recebendo mais atenção da literatura, por este motivo a seção 2.4.2 apresenta especificamente este método.

2.4.2 Mineração de dados

A mineração de dados tem atraído atenção da indústria da informação e da sociedade como um todo, devido à ampla disponibilidade de trabalhar com quantidades significativas de dados e à necessidade iminente de transformar esses dados em informações úteis e conhecimento. As informações e os conhecimentos adquiridos podem ser usados para aplicações que vão desde a análise de mercado, detecção de fraude, controle de produção e exploração da ciência. A mineração de

dados, portanto, refere-se à extração ou "mineração" do conhecimento de grandes quantidades de dados (HAN, KAMBER, 2006).

Para Fayyad *et al.* (1996 b), a mineração de dados é a aplicação de algoritmos específicos para extração de padrões a partir de dados. Depende-se fortemente de técnicas de aprendizagem de máquina, reconhecimento de padrões e estatísticas para encontrar padrões de dados na MD.

Também Pregibon (1996) revela que a área de mineração de dados surgiu a partir de aplicações da comunidade de aprendizagem de máquina. Para o autor a MD é uma mistura de estatística, inteligência artificial e pesquisa de banco de dados.

As principais tarefas de MD são: classificação, regressão, agrupamento ou *clustering*, sumarização e regras de associação. O emprego de uma tarefa e método é eficaz de acordo com a natureza da aplicação em questão. As seções 2.4.3. e 2.4.4 apresentaram especificamente as tarefas de classificação e regras de associação.

2.4.3 Classificação

A tarefa de classificação corresponde a encontrar uma função que associa um caso a uma classe dentro de diversas classes discretas de classificação, esta tarefa pode ser definida como uma função de aprendizado que mapeia itens de dados em uma das diversas classes pré-definidas, consistindo a classificação em obter um modelo baseado em um conjunto de exemplos que descrevem uma função desconhecida (SANTOS; AZEVEDO, 2005).

2.4.3.1 C4.5

O C4.5 é um algoritmo de classificação, uma versão evoluída do algoritmo ID3 (*Iterative Dichotomizer 3*) que consiste em induzir árvores de cima para baixo – *top-down* (QUINLAN, 1993).

O C4.5 é um sistema de aprendizado de máquina que constrói a árvore de decisão a partir da raiz, selecionando o melhor atributo classificador (maior ganho de

informação) dentre todos os atributos do conjunto de dados. Após a escolha, os dados são separados de acordo com as classes do atributo escolhido, gerando uma subdivisão dos dados para cada descendente na árvore. O algoritmo é aplicado recursivamente a cada descendente, até que algum critério de parada seja atingido, como – por exemplo, a exaustão da análise de todos os atributos previsores.

O J48 é uma implementação em Java do algoritmo C4.5 release 8 – última versão pública. Ele foi desenvolvido para ser incorporado ao Weka. Este método apresenta como resultado uma árvore de decisão com poda que pode ser transformada em um conjunto de regras de decisão.

2.4.3.2 PRISM

Trata-se de um algoritmo de classificação que foi proposto para solucionar as limitações identificadas no algoritmo ID3, os quais são: dificuldade de trabalhar com dados ruidosos e a forma de saída em árvore de decisão, que pode ser incompreensível (aos humanos), difícil de manipular (por humanos e computadores) e torna difícil a apresentação de conclusões (de computadores para humanos), tornando seu uso inadequado para muitos domínios.

O PRISM tem como entrada um conjunto de treinamento inserido em um arquivo de conjuntos ordenados de valores para os atributos, cada conjunto a ser utilizado por uma classificação. As informações sobre os atributos e as classificações são a entrada de um processo separado no início do programa e os resultados são transmitidos como regras individuais para cada uma das classificações indicadas dos atributos descritos.

O PRISM concentra-se em encontrar valores relevantes de atributos, devendo identificar subconjuntos de uma classe específica, com a vantagem de apresentar a saída dos dados sob a forma de regras modulares ao invés de uma árvore de decisão (CENDROWSKA, 1987).

O objetivo desta tarefa é descobrir um relacionamento entre um atributo meta e um conjunto de atributos previsores, de modo que o processo de classificação possa usar esse relacionamento para prever a classe de um exemplo novo e não conhecido. Por meio da classificação é possível determinar o valor de

um atributo a partir dos valores de um subconjunto dos demais atributos da base de dados.

Uma das formas de teste da base que está sendo executada a tarefa ou do algoritmo é a validação cruzada com k-partições (*k-folds*) sendo a base de dados dividida, de forma aleatória, em k partições. Na sequência, uma das partições é selecionada para compor o conjunto de teste e os k-1 restantes são usados para compor o conjunto de treinamento. Este procedimento é repetido até que todas as k partições sejam selecionadas como conjunto de teste. O resultado final é fornecido pela média dos resultados obtidos em cada partição (KOHAVI, 1995).

Segundo Cendrowska (1987), para cada classe δ_n , deve-se:

- 1) calcular a probabilidade $p(\delta_n | \alpha_x)$ de ocorrências de δ_n para cada par de atributo-valor α_x ;
- 2) selecionar o α_x para cada $p(\delta_n | \alpha_x)$ com maior ganho de informação e criar um subconjunto de treinamento englobando as instâncias que contêm α_x selecionado;
- 3) repetir os passos 1 e 2 para este subconjunto até ele conter somente instâncias da classe δ_n . A regra induzida é a conjunção de todos os atributos usados na criação do subconjunto homogêneo de dados;
- 4) remover todas as instâncias cobertas por esta regra do conjunto de treinamento;
- 5) repetir os passos 1 a 4 para o subconjunto formado no passo 4, até todas as instâncias da classe δ_n serem removidas.

Ao final o conjunto de treinamento é restaurado e retorna-se ao passo 1, para execução da próxima classe.

2.4.1 Associação – Algoritmo Apriori

O objetivo desta tarefa é identificar associações não visíveis entre registros de dados que estão ou devem estar relacionados, tal que a presença de alguns deles em uma transação implique na presença de outros na mesma transação, descobrindo-se dessa forma uma relação já consolidada ou no mínimo uma

tendência. O objetivo, então, é encontrar associações relevantes entre os itens, do tipo X (antecedente) $\rightarrow Y$ (consequente). Como em toda tarefa de MD existem métodos que as implementam e no caso da tarefa de associação, o método mais conhecido e utilizado é o de Regras de Associação. Uma regra de associação caracteriza o quando a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (AGRAWAL; SRIKANT 1994).

A frequência dos elementos no banco de dados de uma regra de associação, são o fator de suporte e de confiança. O fator de confiança é o grau para o qual a regra é verdadeira frente aos registros individuais. É calculado da seguinte forma: número de transações que suportam a regra, dividido pelo número de transações que suportam somente o corpo da regra (BERRY; LINOFF, 1997). A confiança é uma medida da força das regras descobertas, vale dizer, a percentagem de itens de X que possuem Y .

O fator de suporte se refere a quantidade de ocorrências onde um determinado item ou conjunto de itens aparecem em transações distintas relacionadas com o número total de transações operadas, que é calculado da seguinte forma: ocorrência do item ou conjunto de itens dividido pelo número de transações operadas (BERRY; LINOFF, 1997). O suporte pode ser interpretado como a frequência com que ocorrem os padrões em toda a base; é a percentagem dos itens da base que possuem X e Y .

Um dos algoritmos existente para gerar regras de associação é o Apriori, que segundo Agrawal *et al.* (1996) realiza buscas sucessivas em toda a base de dados, encontrando todos os conjuntos de itens que são frequentes, denominados *itemsets frequentes* (L_k). O Apriori faz uso de duas funções: a função *Apriori_gen*, para gerar os candidatos e eliminar aqueles que não são frequentes, e a função *Genrules*, utilizada para extrair as regras de associação.

A respeito de seu funcionamento, primeiramente é realizada a contagem de ocorrências dos itens para determinar os *itemsets frequentes* de tamanho unitário (1-*itemsets frequentes*). Os passos posteriores, k , consistem de duas fases. Primeiro, os *itemsets frequentes* L_{k-1} , encontrados no passo anterior ($k-1$) são utilizados para gerar os conjuntos de itens potencialmente frequentes, os *itemsets candidatos* (C_k). Na sequência, é realizada uma nova busca no banco de dados, contando-se o suporte de cada candidato em C_k .

Os *itemsets candidatos* são gerados inicialmente tomando como argumento L_{k-1} , o conjunto de todos $(k-1)$ -*itemsets frequentes*. Para isto, utiliza-se a função *Apriori_gen*, que retorna um superconjunto de todos os k -*itemsets frequentes*. A intuição por trás desse procedimento é que se um *itemset* X tem suporte mínimo, todos os seus subconjuntos também terão.

O último passo é a descoberta das regras de associação, obtida por meio da função *Genrules*. A geração de regras, para qualquer *itemset frequente*, significa encontrar todos os *subsets* não vazios de I . Desta forma, para todo e qualquer *subset* a , produz-se uma regra $a \rightarrow (I - a)$ somente se a razão ($\text{suporte}(I)/\text{suporte}(a)$) é ao menos igual a confiança mínima estabelecida pelo usuário. Para gerar regras com múltiplos consequentes, são considerados todos os *subsets*. Por exemplo, dado um *itemset* $ABCD$, considera-se primeiro o *subset* ABC , seguido de AB , etc. Se $ABC \rightarrow D$ não atinge uma confiança suficiente ($\text{confiança} < \text{minconf}$), não é necessário verificar se $AB \rightarrow CD$.

2.5 TRABALHOS RELACIONADOS

Neste tópico será feita uma síntese dos trabalhos que contribuíram com o desenvolvimento desta pesquisa.

Digiampietri *et al.* (2012) identificam os desafios para utilizar os currículos da Plataforma Lattes, podendo ser atribuídos à quantidade de dados, ao preenchimento manual, além do uso de dados semiestruturados. Este trabalho apresenta um banco de dados desenvolvido a partir da mineração dos dados de mais de um milhão de Currículos Lattes. Os dados da Plataforma Lattes utilizados no artigo embasaram a etapa de coleta de dados (seção 3.3) do modelo proposto neste trabalho.

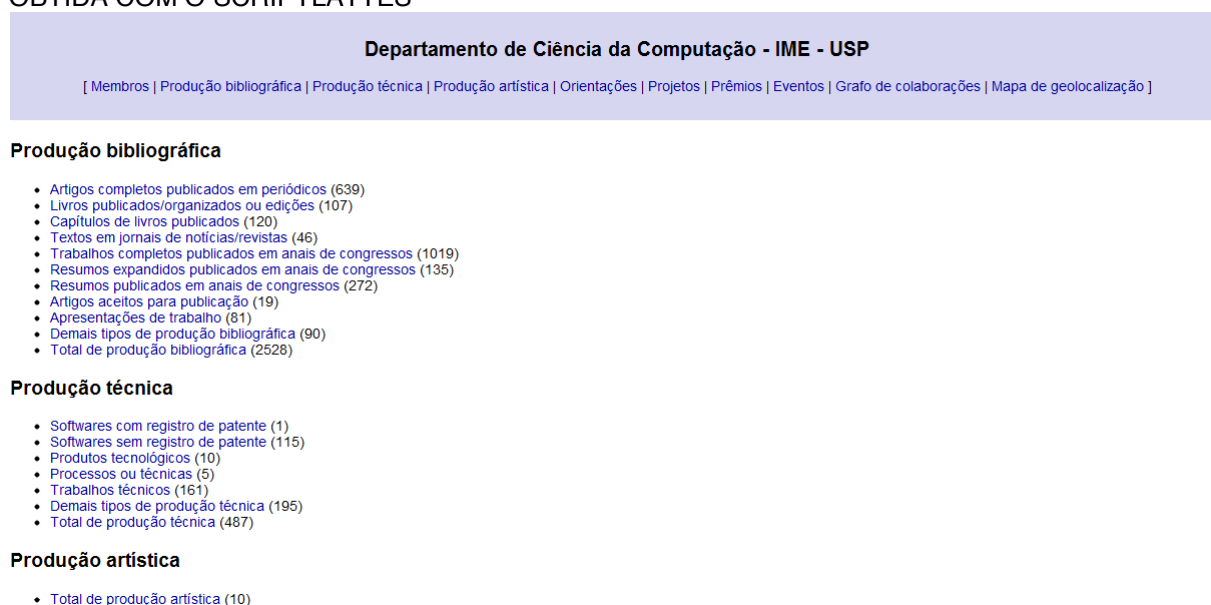
O sistema de avaliação e apoio institucional da Universidade de São Paulo (USP), que recebeu o nome de Tycho, foi desenvolvido para atender a necessidade de relatórios de avaliações administrativas e acadêmicas, tendo como princípio integrar as atividades de coleta de dados, avaliação e planejamento da USP a fim de integrar indicadores para auxiliar em atividades como avaliações e gestão acadêmica (USP, 2013). Conforme Mugnaini *et al.* (2012), este sistema tem como fonte de dados: sistemas centrais mantidos pelo Departamento de Informática da

Reitoria da USP, currículo Lattes e Grupos de Pesquisa do CNPq. Entre os recursos fornecidos pelo sistema estão os grafos de colaboração, indicadores de produção bibliográfica, produção técnica, produção artística, entre outros, a partir do ano de 1996.

No trabalho de Mena-Chalco e César Jr. (2009) é apresentado o software de código aberto scriptLattes, que possibilita a geração de relatórios acadêmicos de grupos com base em currículos da base de dados Lattes. O sistema é composto por seis módulos que permitem a geração de grafos de colaboração entre os membros dos grupos e a criação de relatórios de produção bibliográfica, técnica e artísticas, além de orientações acadêmicas.

A Figura 17 apresenta o módulo de produção do scriptLattes aplicado ao departamento de ciência da computação da Universidade de São Paulo.

FIGURA 17 – EXEMPLO DO RELATÓRIO DE PRODUÇÃO BIBLIOGRÁFICA, TÉCNICA E ARTÍSTICA OBTIDA COM O SCRIPTLATTES

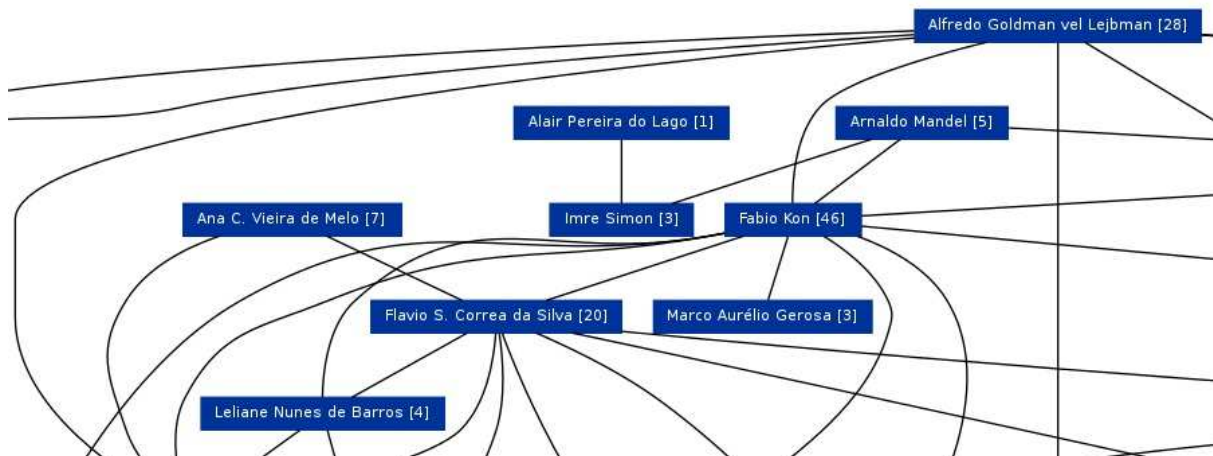


FONTE: Mena-Chalco; César Jr. (2012)

Na Figura 17 é possível identificar os tipos de produção bibliográfica, técnica e artística que os membros do departamento têm, sendo que ao lado de cada tipo de produção consta o número total de publicações daquele determinado tipo, ao escolher uma das opções o usuário pode visualizar a referência de cada documento separado por ano de publicação.

Já na Figura 18 identifica-se parte do grafo de colaboração do grupo. Sendo que, “para cada nó o valor entre colchetes indica o número de produções feitas em colaboração apenas com os outros membros” (MENA-CHALCO; CÉSAR JR., 2012).

FIGURA 18 – EXEMPLO DE PARTE DO GRAFO DE COLABORAÇÃO GERADO COM O SCRIPTLATTES



FONTE: Mena-Chalco; César Jr. (2012)

A pesquisa de Ferreira (2007) consiste no desenvolvimento e análise de um modelo para mapeamento do conhecimento. A ferramenta desenvolvida buscou representar graficamente o conhecimento dos indivíduos. O mapa de conhecimento conforme proposto por este estudo, foi aplicado em uma empresa de desenvolvimento, suporte e comercialização de softwares de gestão de hospitais, clínicas e restaurantes, especificamente na área de suporte ao cliente do produto para grandes hospitais. As evidências empíricas da pesquisa apontaram que o modelo proposto, que se baseou nos diagramas de relevância desenvolvido por Howard (1989) era aplicável, com baixo nível de complexidade, mas com grande demanda de tempo.

Le-Khac *et al.* (2007) apresentam um “mapa do conhecimento” que é a representação do conhecimento extraído, visando facilitar a visualização e a avaliação dos resultados de mineração de dados. Abordam a questão da grande quantidade de dados coletados em diversos campos o que gera o problema da falta de eficiência na gestão do conhecimento em mineração de dados distribuídos (*Distributed Data Mining - DDM*), as técnicas existentes de DDM analisam parcialmente os dados locais e então geram modelos globais de agregação dos resultados locais. No entanto, como estas etapas são dependentes podem produzir modelos globais de dados incorretos ou ambíguos.

O trabalho de Pirola (2002) apresenta uma proposta de desenvolvimento de uma ferramenta computacional para apoiar o mapeamento do conhecimento organizacional, neste trabalho é seguido o modelo de Probst, Raub e Romhardt (2002) e Cavalcanti, Gomes e Pereira (2001) para a metodologia, sendo apresentado à arquitetura tecnológica necessária para o funcionamento da ferramenta. Como resultados demonstra-se como foram implementadas as características propostas para a metodologia na ferramenta desenvolvida.

O Quadro 4 mostra um mapeamento dos trabalhos relacionados apresentados, contendo as características relevantes para esta pesquisa.

QUADRO 4 – CARACTERÍSTICAS DOS TRABALHOS CO-RELATOS

Autor(es)	Proposta de mapa do conhecimento	Uso de grafo/árvore	Uso da Plataforma Lattes	Uso de mineração de dados
Digiampietri <i>et al.</i> (2012)	NÃO	SIM	SIM	SIM
USP (2012) / Mugnaini <i>et al.</i> (2012)	NÃO	SIM	SIM	NÃO
Mena-Chalco e César Jr. (2009)	NÃO	SIM	SIM	NÃO
Ferreira (2007)	SIM	NÃO	NÃO	NÃO
Le-Khac <i>et al.</i> (2007)	SIM	SIM	NÃO	SIM
Pirola (2002)	SIM	NÃO	NÃO	NÃO

FONTE: A Autora (2012)

A seguir a seção 3 apresenta a metodologia empregada para esta pesquisa, bem como seu ambiente, base de dados e materiais utilizados.

3 METODOLOGIA DE PESQUISA

Neste trabalho, propõe-se um modelo para construção de mapas de conhecimento para instituições, aplicando-o em uma instituição acadêmica. A presente pesquisa de investigação qualitativa caracteriza-se pela abordagem descritiva, tendo como foco a gestão do conhecimento, apresentando a criação de um modelo capaz de resgatar informações de pesquisadores. O enfoque da proposta apresentada está na definição do fluxo de atividades para geração do mapa e nas alternativas disponíveis para sua visualização. As seções de 3.1 a 3.3 descrevem o ambiente de pesquisa no qual a proposta é aplicada e sua limitação, a seção 3.4 apresenta os materiais utilizados para possibilitar a implantação da pesquisa.

3.1 O AMBIENTE DE PESQUISA

Para a construção da prova de conceito do modelo proposto e avaliação da viabilidade da sua implantação, utilizou-se neste trabalho como ambiente de pesquisa a instituição Universidade Federal do Paraná. Dentro deste universo, limitou-se a construção do mapa do conhecimento (MC) aos professores/pesquisadores dos programas de pós-graduação de mestrado e doutorado, *stricto sensu*.

A Universidade Federal do Paraná - UFPR é a mais antiga universidade brasileira, criada em 19 de dezembro de 1912 e iniciando seu funcionamento em 1913. Em 2012, segundo dados da Pró-Reitoria de Pesquisa e Pós-Graduação – PRPPG, a UFPR contava com sessenta e dois programas de mestrados acadêmicos, quarenta programas de doutorados e quatro mestrados profissionalizantes (PRPPG, 2012a). Somente nos últimos 10 anos foram criados trinta e três novos Programas de Pós-Graduação. Na última avaliação trienal da CAPES, de 2007 a 2009, 43 programas dos 106 avaliados obtiveram conceito de 5 e 4 enquanto 20 programas obtiveram conceito 3 (PRPPG, 2012b). A nota atribuída pela CAPES obedece a uma escala de 1 a 7.

Quanto à qualificação e capacitação dos professores/pesquisadores a UFPR tem realizado investimentos por meio de uma política de qualificação, que resultou em um aumento do quadro permanente de doutores de 37,9% em 2001 para 66,7% em 2011, e na redução do quadro de mestres de 35,0% para 21,3%, no mesmo período. O resultado desta política também foi positivo em relação à produção científica, já que o número de publicações em periódicos indexados internacionalmente na base de dados da *Web of Science* do *Institute for Scientific Information* (ISI), aumentou no último triênio em 66,6%, totalizando 2.650 artigos e 17.999 citações. Esses totais representam 47% das citações de trabalhos publicados por pesquisadores da UFPR entre os anos de 1945 e 2009 (PRPPG, 2012b).

3.2 COLETA DOS DADOS

Para a realização da pesquisa foram coletados dados existentes na Plataforma Lattes dos professores/pesquisadores vinculados aos programas de pós-graduação *stricto sensu* da UFPR. Estes dados são utilizados para apresentação do modelo de mapa do conhecimento proposto. A partir dos dados disponíveis na base de dados, são consideradas as informações a seguir:

- 1) áreas de interesse: abrange a: Formação Acadêmica e Áreas de Atuação, as quais também podem ser definidas como Áreas do Conhecimento do pesquisador. Técnicas de mineração de dados foram aplicadas nas áreas de interesse visando a descoberta de relacionamentos entre as áreas de pesquisa. Pretende-se assim trazer como resultado de uma pesquisa do usuário do mapa do conhecimento, não só a área pesquisada, mas também as áreas a ela relacionadas;
- 2) produção científica: composta por: artigos publicados em periódicos/revistas científicas, capítulos de livros, livros. Quando se trata de capítulo de livro é considerado o título do capítulo, desconsiderando o título do livro. Considera-se que o título de cada documento, publicado no mesmo ano é de caráter único e por esta forma é possível identificar as colaboradores entre autores;

- 3) palavras-chave: a fim de auxiliar a identificação da produção científica também realizou-se o levantamento de todas as palavras-chave cadastradas na base, referente às publicações científicas apontadas no item anterior;
- 4) dados pessoais: são utilizadas apenas as informações referentes ao local de trabalho, como: departamento e setor vinculados aos professores/pesquisadores. O nome do pesquisador bem como seu número de telefone e e-mail não foram coletados, somente o número identificador do arquivo XML é apresentado.

Para viabilizar a composição da base de dados deste estudo, são utilizados arquivos XML da base da Plataforma Lattes mantida pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), em que cada pessoa é responsável por cadastrar seu currículo Lattes e manter o mesmo atualizado, sendo esta plataforma de acesso público pela internet. A presente pesquisa tem como área de estudo os programas de pós-graduação *stricto sensu* da UFPR, utilizando os arquivos da plataforma cedidos pela Pró-Reitoria de Pesquisa e Pós-Graduação desta instituição dos professores/pesquisadores internos que fazem parte destes programas. Estes arquivos compõem uma base com dados cadastrados na Plataforma Lattes, contendo as mesmas informações de acesso público pela internet, como formação acadêmica e produção científica de todos os professores/pesquisadores com atualização até às onze horas da manhã do dia vinte e dois de dezembro de 2011, por ser esta a data em que os arquivos XML foram coletados.

Os arquivos XML foram disponibilizados à PRPPG devido a um convênio entre a UFPR e o CNPq e que toda instituição de pesquisa e inovação no Brasil pode solicitar, por meio de encaminhamento de ofício à presidência da CNPq. Desta forma, as instituições que possuem o protocolo de cooperação técnica firmado têm garantido o direito de acesso à extração de dados da Plataforma Lattes, de seus pesquisadores, professores e alunos, em formato XML (PLATAFORMA LATTES, 2013).

Na versão da Plataforma Lattes de janeiro de 2013, o formato de arquivo XML do Currículo Lattes também estava disponível a todos os usuários da plataforma. Ao visualizar um currículo uma das opções apresentadas permite baixar o currículo que está sendo visualizado em XML (Figura 19).

FIGURA 19 – OPÇÃO PARA BAIXAR CURRÍCULO LATTES EM ARQUIVO FORMATO XML



FONTE: Plataforma Lattes (2013)

3.3 UNIVERSO E AMOSTRAGEM

O universo deste estudo compreende os docentes da Universidade Federal do Paraná. Conforme dados da PROGEPE *apud* PROPLAN (2011) são 2.168 docentes e mais 163 docentes substitutos ou visitantes. Tendo como amostra os docentes internos vinculados aos programas de pós-graduação *stricto sensu* desta universidade que totalizam segundo dados passados pela PRPPG (2011) 1.073 docentes, portanto a palavra pesquisador/professor, que ora pode ser utilizada apenas como pesquisador neste trabalho compreende apenas os docentes desta universidade especificamente dos programas *stricto sensu*, não abrangendo os mestrandos e doutorandos. A escolha desta instituição motivou-se, pois, além de ser um ambiente voltado à educação e à pesquisa, investe na qualificação e capacitação dos professores/pesquisadores, o qual o compartilhamento do conhecimento que é promovido pelo mapeamento do conhecimento pode contribuir. Ademais, o fator de disponibilidade de acesso aos dados influenciou nesta escolha, tendo em vista o convênio que a UFPR tem com a CNPq que permite a extração dos dados da Plataforma Lattes.

Como a pesquisa é de caráter qualitativo não requer elevado nível de precisão e rigor estatístico. A amostra do estudo exploratório foi selecionada pelo critério de intencionalidade e de acessibilidade. Conforme Gil (1994, p.96), “uma amostra intencional, em que os indivíduos são selecionados a partir de certas características tidas como relevantes pelos pesquisadores e participantes, mostra-se mais adequada para a obtenção de dados de natureza qualitativa”.

3.4 MATERIAIS

As seções 3.4.1 a 3.4.10 apresentam as ferramentas computacionais utilizadas para a construção desta pesquisa.

3.4.1 Plataforma Lattes

A Plataforma Lattes é uma arquitetura de informações desenvolvida para o CNPq gerir atividades de planejamento, gestão e operacionalização de fomento do próprio CNPq e também de outras agências de fomento federais e estaduais, fundações estaduais de apoio à ciência e tecnologia, instituições de ensino superior e institutos de pesquisa.

A Plataforma Lattes se tornou estratégica não só para as atividades de planejamento e gestão, como também para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.

A disponibilização dos dados de forma pública da Plataforma na internet garante maior transparência e confiabilidade às atividades de fomento do CNPq e das agências que a utilizam, sendo uma fonte inesgotável de informações para pesquisas (PLATAFORMA LATTES, 2013).

3.4.2 *Extensible Markup Language* - XML

O XML é um formato de arquivo de texto com uma linguagem de marcação extensível de dados cujo principal objetivo é o intercâmbio de informações, foi desenvolvido para enfrentar os desafios da grande escala de publicação eletrônica e para contribuir na troca da ampla variedade de dados.

XML foi desenvolvido por um Grupo de Trabalho XML (conhecido como o Conselho Editorial SGML Review) formado sob os auspícios da World Wide Web Consortium (W3C) no ano de 1996 e presidida por Jon Bosak da Sun Microsystems. Teve a participação ativa de um Grupo de Interesse Especial XML (anteriormente

conhecido como o Grupo de Trabalho SGML), também organizado pela W3C (W3C, 2013).

3.4.3 PostgreSQL

O PostgreSQL é um sistema gerenciador de banco de dados objeto-relacional de código aberto, com mais de 15 anos de desenvolvimento ativo e provê uma arquitetura confiável, que confere integridade dos dados e corretude.

Este banco de dados pode ser executado na grande maioria dos sistemas operacionais, incluindo Linux (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) e Windows. Tem suporte completo para chaves estrangeiras, associações, visões, gatilhos e *store procedures* (em múltiplas linguagens). Inclui suporte para a maioria dos tipos definidos no padrão SQL:2008, incluindo INTEGER, NUMERIC, BOOLEAN, CHAR, VARCHAR, DATE, INTERVAL e TIMESTAMP. Também possui suporte para armazenamento no formato binário de grandes objetos como imagens, sons ou vídeos. Possui interface nativa para programação C/C++, Java, .Net, Python, Ruby, Tcl, ODBC entre outros e uma documentação completa (POSTGRESQL, 2013).

3.4.4 Python

Python é uma linguagem de programação de uso e distribuição livre que utiliza padrões da programação orientada a objeto. Esta linguagem é utilizada em uma grande variedade de aplicações, sendo comparada muitas vezes a outras linguagens como: Java, Perl e Ruby. Tem como características: ter uma sintaxe clara, orientação a objetos intuitiva, extensas bibliotecas padrão e módulos para praticamente todas as tarefas. Muitos distribuidores UNIX já vem com o Python instalado em seus computadores e também em alguns computadores que possuem o sistema operacional Microsoft. O *Python Software Foundation* (PSF) é uma organização sem fins lucrativos que detém os direitos de propriedade intelectual da linguagem de programação Python (PYTHON, 2012).

3.4.5 Django

O Django é um framework web desenvolvido em linguagem Python para tornar tarefas comuns de desenvolvimento de software para internet mais rápidas e fáceis. Este framework foi projetado para lidar com os desafios de prazos curtos de uma redação e os requisitos rígidos dos desenvolvedores web que o escreveram. O Django possibilita a construção de aplicações com recursos administrativos (alterar, incluir ou excluir dados) que podem ser gerados automaticamente (DJANGO, 2013).

3.4.6 Java

Java é uma linguagem de programação orientada a objetos desenvolvida em 1991, por um grupo de engenheiros da empresa Sun Microsystems, chamado "*Green Team*", liderado por James Gosling. Em 1995 o navegador de internet Netscape Navigator incorporou a tecnologia Java que atualmente é utilizada nas mais diversas aplicações em redes, como jogos, telefones móveis, softwares empresariais (ORACLE, 2013).

3.4.7 Xstream

Xstream é uma biblioteca Java de código aberto para criar objetos de arquivos XML para Java e vice-versa, como vantagens da utilização podem ser citados: nenhuma informação é duplicada do XML, ocorrendo alguma exceção devido a XML com problemas de formatação são fornecidos mensagens de erros detalhadas para auxiliar na resolução de problemas. Além disto, esta biblioteca é capaz de gerar representações XML em praticamente qualquer objeto Java sem implicar em modificações nas classes originais (XSTREAM, 2013).

3.4.8 JavaScript InfoVis Toolkit – JIT

O JIT é uma biblioteca Java que dispõe de ferramentas para criação de formas interativas da visualização de dados para web, como diversos tipos de árvores (JIT, 2013).

3.4.9 Microsoft Office Excel

O Microsoft Office Excel, também chamado de Excel é uma planilha eletrônica, que faz parte de um pacote de aplicativos da Microsoft para criar documentos, apresentações, gerenciar e-mail, planilhas. Entre as facilidades deste software está a criação de tabelas, gráficos, realização de cálculos. O Excel pode ser instalado em computadores com sistema operacional Microsoft e Mac OS, a versão mais recente do programa é a 2010 para Windows e 2011 para Mac OS, sendo que sua primeira versão foi lançada em 1987 (MICROSOFT, 2013).

3.4.10 Weka

Weka é um software de fonte aberta de mineração de dados desenvolvido na linguagem Java pela Universidade de Waikato, na Nova Zelândia. Este software possui uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados, sendo que estes algoritmos podem ser aplicados diretamente a uma base de dados ou chamado a partir do próprio código Java. Nesta pesquisa os algoritmos foram aplicados à base de dados, utilizando a versão 3.6.4 do software. O Weka possui algoritmos de classificação, regressão, agrupamento e regras de associação (HALL *et al.*, 2013).

A seção 4 apresenta como o modelo de mapa do conhecimento proposto foi construído, aplicação para a base de dados utilizada a resultados obtidos com a caracterização e mineração de dados.

4 CONSTRUÇÃO DO MODELO

O modelo proposto nesta pesquisa é formado por três elementos: componentes de entrada, processos e produtos que serão descritos nas seções 4.1 a 4.5. Sua construção tem como base os modelos apresentados por Probst, Raub e Romhardt (2002, p. 69) e Ebener et al. (2006). Referente a estrutura dos dados a seção 4.4 apresenta o modelo entidade-relacionamento proposto para o banco de dados.

A construção do modelo se divide em cinco etapas, que tem como base a literatura. O Quadro 5 apresenta as etapas do modelo proposto neste trabalho para a construção do mapa do conhecimento, comparando com as propostas com os dois modelos da literatura constantes no quadro.

QUADRO 5 - COMPARAÇÃO ENTRE OS MODELOS DA LITERATURA E O MODELO PROPOSTO

Etapas	Modelo proposto por Probst, Raub e Romhardt (2002, p. 69)	Modelo apresentado por Ebener et al. (2006)	Modelo proposto neste trabalho
1	Identificar processos intensivos em conhecimento.	Dados brutos são adquiridos de uma ou mais fontes utilizando pesquisas.	Definição do significado dos componentes de entrada de acordo com o objetivo da elaboração do mapa do conhecimento.
2	Localizar ativos do conhecimento e detentores de conhecimento que são relevantes.	Dados brutos são manipulados por análises básicas para produzir dados de primeira ordem que são adequados para gerar mapas de conhecimento.	Importação dos dados originais para a base de dados do mapa do conhecimento, mapeando os atributos de origem para os componentes: ator, grupo de trabalho e artefato produzido.
3	Indexar os ativos e os detentores.	Os dados de primeira ordem são armazenados em uma base de dados central, a qual é frequentemente conhecida como base de dados de mapa do conhecimento do inglês <i>knowledge mapping database</i> .	Disponibilização dos resultados do processo anterior para acesso e consulta simples.
4	Integrar os ativos indexados em um sistema de navegação que esteja ligado ao processo.	Dados brutos armazenados na base de dados são processados por análises, agregações e contextualizações e resultam em dados de ordem superior.	Processamento sob demanda dos dados armazenados no banco de dados gerando visualizações alternativas das informações.

Étapas	Modelo proposto por Probst, Raub e Romhardt (2002, p. 69)	Modelo apresentado por Ebener <i>et al.</i> (2006)	Modelo proposto neste trabalho
5	Possibilitar mecanismos descentralizados de atualização.	Por meio da visualização dos dados de primeira ordem e ordem superior, os mapas de conhecimento podem ser gerados, permitindo acesso ao conhecimento disponível na organização ou em um domínio específico do trabalho.	Disponibilização para que o usuário que tem acesso a parte administrativa do sistema possa realizar edição(ões) e atualização(ões) das informações importadas para o banco de forma manual por meio da interface web.

FONTE: A Autora (2012)

Pode-se destacar como vantagens do modelo proposto:

- pode ser aplicado em diversas instituições e plataformas, tendo em vista que, os componentes de entrada do modelo (grupos de trabalhos, atores e artefatos produzidos) podem adequar-se a diferentes cenários presentes nas instituições;
- o modelo expõem eventuais problemas associados aos dados fornecidos, como: sinonímias, erros de preenchimento e ausências de dados;
- organização dos dados da instituição, permitindo pesquisas por diferentes critérios e fazendo uso de técnicas de visualização de informações para apresentação dos resultados;
- enfatizar a visualização do relacionamento entre os componentes e a navegação por meio dos relacionamentos existentes entre eles.

As características descritas, também apoiaram a diferenciação desta pesquisa com outras relativas a modelos de mapa de conhecimento que foram analisados. Por não encontrar estes atributos em outros MC, identificou-se assim a necessidade de proposição de um novo modelo.

4.1 COMPONENTES

Os componentes de entrada são as informações necessárias para a construção do mapa do conhecimento. Elas estão classificadas em três categorias:

atores, grupos de trabalho e artefatos produzidos. Estas categorias foram assim definidas com base na nomenclatura adotada no padrão UML (*Unified Modeling Language*) que é uma linguagem utilizada para modelagem de sistemas de software. O termo ator se refere ao papel que um usuário desempenha com relação a um sistema e artefato normalmente são arquivos (FOWLER, 2005). Já a categoria: grupos de trabalhos, recebeu este termo por representar as relações entre os atores.

As categorias foram definidas de forma a não estarem associadas a uma base de dados específica como a Plataforma Lattes, permitindo assim que o modelo possa ser aplicado a diferentes bases de dados.

No contexto deste trabalho, os **atores** são os componentes responsáveis por produzir artefatos e envolverem-se em grupos de trabalho. Os atores podem ser funcionários, pesquisadores, alunos ou até elementos externos à instituição.

Os **grupos de trabalho** são utilizados para representar associações entre atores. O significado real de um grupo de trabalho depende do objetivo definido no início da construção do mapa e pode significar, por exemplo: o departamento da instituição, laboratório, projeto, evento como congresso ou tema de interesse em comum. Alguns grupos de trabalho podem ser determinados de maneira simples, baseado em consultas às informações disponíveis, como departamento ao qual o ator está associado, outros podem demandar uso de técnicas de mineração de dados a fim de descobrir padrões de associação ou relacionamentos. Este modelo não fixa as técnicas a serem utilizadas para definição dos grupos de trabalho, permitindo desta forma que múltiplas técnicas possam ser utilizadas na identificação dos grupos de trabalho.

Os **artefatos** produzidos, no escopo deste trabalho, estão definidos como todo resultado de uma atividade do ator, seja individualmente, como uma dissertação de mestrado, seja por um grupo de trabalho, como em uma documentação de projeto com coautores, ou artigo com coautores. Uma das vantagens em utilizar artefatos produzidos no mapa de conhecimento é permitir a visualização das informações associadas, não somente a formação e interesses do ator, mas também as informações associadas a sua produção intelectual.

4.2 PROCESSOS

O modelo proposto possui dois processos associados. O primeiro processo é responsável por importar os dados da fonte original para o modelo de dados do método. Neste processo, os atributos originais são mapeados para os componentes de entrada: ator, grupo de trabalho ou artefato produzido. O resultado deste processo é o armazenamento em base dos dados que são de interesse para a elaboração do mapa do conhecimento. Estes dados são armazenados utilizando o modelo proposto no método e ficam disponíveis para consulta imediata ou para serem utilizados no segundo processo proposto no método.

O segundo processo é responsável por fornecer ferramentas para que o usuário possa consultar o mapa do conhecimento e visualizar os resultados com diferentes técnicas como páginas amarelas e grafos. Este processo pode ser dividido em sub-processos, cada qual responsável por gerar uma forma de visualização das informações. Ao contrário do processo anterior, o segundo processo é interativo e elabora os produtos de acordo com as solicitações recebidas, permitindo que os produtos sejam personalizados de acordo com o interesse do usuário do mapa do conhecimento.

4.3 PRODUTOS

O modelo possui dois produtos, no contexto deste trabalho os produtos são os resultados dos processos. O primeiro produto consiste nos dados importados para a base de dados do mapa do conhecimento e disponíveis para consultas simples. Consultas simples no escopo deste trabalho significam consultas baseadas em atributos dos componentes, como por exemplo: nome do ator, área de interesse do ator, nome do grupo de trabalho, ano de publicação do artefato. Além das consultas simples, os dados são visualizados utilizando listas ou gráficos de linha ou coluna.

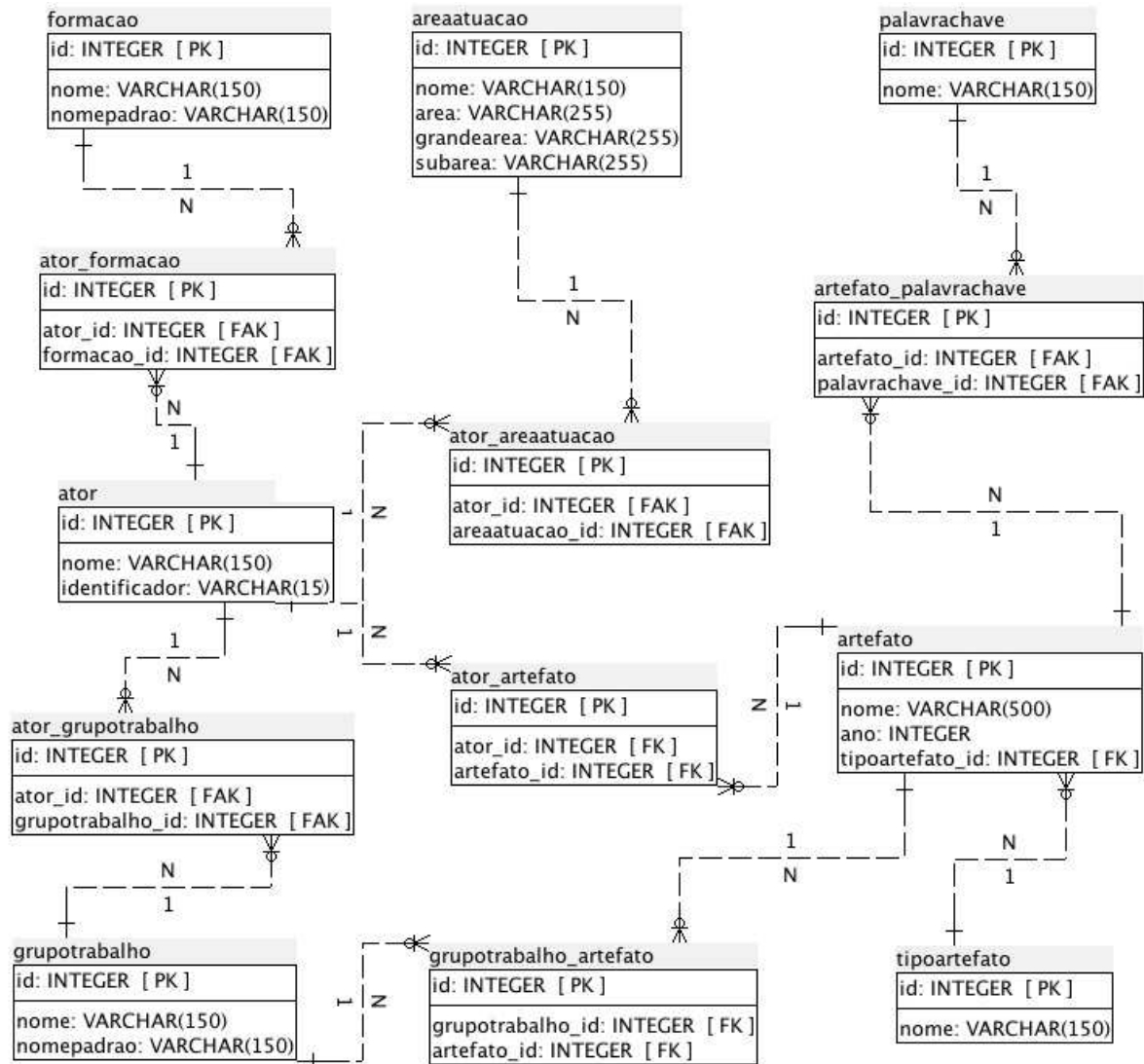
O segundo produto é composto por técnicas de visualização mais elaboradas como visualização por meio de grafos. Além disto, a informação apresentada pode ser o resultado de uma técnica de mineração de dados.

4.4 MODELO ENTIDADE-RELACIONAMENTO DO BANCO DE DADOS

A implementação da prova de conceito do modelo de construção de mapa do conhecimento faz uso de um banco de dados relacional para armazenamento dos dados obtidos das fontes de dados originais. O modelo entidade-relacionamento - MER utilizado no banco de dados reflete a estrutura e relacionamentos existentes entre atores, grupos de trabalho, formações, áreas de atuação, artefatos e palavras-chave, além de manter informações auxiliares, conforme Figura 20.

Os dados referentes aos componentes de entrada são armazenados nas tabelas, conforme Figura 20.

FIGURA 20 - MODELO ENTIDADE-RELACIONAMENTO



FONTE: A Autora (2012)

As tabelas do modelo são descritas no Quadro 6.

QUADRO 6 - DESCRIÇÃO DAS TABELAS DO MER

Nome da tabela	Descrição
Ator	tabela contendo a identificação única do funcionário/colaborador e o nome deste
formacao	tabela contendo o identificador único da formação, podendo ser o curso técnico, ou a graduação, ou seja o dado mais adequado para a instituição
grupotrabalho	tabela contendo o identificador único da associação entre os atores e o nome que foi atribuída a esta associação

Nome da tabela	Descrição
artefato	tabela contendo o identificador único referente ao artefato e o tipo de artefato, por exemplo, documento técnico produzido
tipoartefato	tabela contendo o identificador único referente ao tipo de artefato, nome do tipo de artefato, por exemplo o título, neste caso o tipo pode ser: manuais, procedimentos, documentação de projetos, produção científica
palavrachave	tabela contendo o identificador único referente a palavra-chave e o atributo nome que se refere a palavra-chave propriamente dita

FONTE: A Autora (2012)

Ainda na Figura 20, têm-se as tabelas auxiliares do MER, sendo elas: ator_formacao, ator_areaatuacao, ator_grupotrabalho, ator_areaatuacao, ator_artefato e artefato_palavrachave.

Estas tabelas servem para representar modelos de relacionamento NxN. Por exemplo, a tabela auxiliar ator_formacao representa que um ator pode ter uma ou mais formações, como graduações, sendo que, cada curso de graduação pode ter vários atores.

Outro relacionamento NxN neste modelo é que um ator pode ter um ou mais artefatos, e um artefato pode estar associado a um ou mais atores. Um exemplo prático desta situação é um artigo com dois ou mais autores. No modelo proposto, evita-se cadastrar um mesmo artigo em duplicidade, a fim de evitar redundância na base de dados. Entende-se aqui que um artefato com o mesmo título e mesmo ano de publicação é único. Assim, uma vez cadastrado o artigo, o mesmo pode ser associado a vários atores. A seção 5.1 explica como foi aplicado esta característica à prova de conceito.

Um artefato pode ter associado uma ou mais palavras-chave e possui uma relação 1xN com tipoartefato, ou seja, um artefato está associado a somente um tipo de artefato, já um tipo de artefato pode estar associado a um ou mais artefatos. Um artefato pode estar associado a um grupo de trabalho, ou seja, um grupo de trabalho pode produzir um artefato no decorrer das suas atividades.

Um grupo de trabalho também está associado aos atores. O grupo de trabalho pode ser interpretado de diferentes maneiras de acordo com o objetivo do mapa do conhecimento e o ambiente que ele está sendo aplicado. Ele pode ser o

departamento ao qual o ator está associado, pode ser um projeto de pesquisa, um projeto comercial, pode ser um evento ou um conjunto de autores que desenvolveram um manual.

4.5 CARACTERIZAÇÃO DO BANCO DE DADOS

Os 1.073 currículos foram coletados em 20 de dezembro de 2011, seguindo uma relação de professores/pesquisadores disponibilizada pela PRPPG.

Os dados considerados na base foram:

- a) identificador;
- b) departamento;
- c) formação;
- d) área de atuação;
- e) artigos publicados em periódicos (título e suas respectivas palavras-chave);
- f) capítulos de livros (título do capítulo e suas respectivas palavras-chave);
- g) livros (título e suas respectivas palavras-chave).

Com exceção do identificador que é um número único gerado pela própria base, que é utilizado em detrimento ao nome do pesquisador, todos os dados foram fornecidos e atualizados pelos próprios pesquisadores. A área de atuação, especificamente a grande área e a área são os únicos campos que contam com vocabulário controlado, ou seja, que disponibilizam ao usuário uma lista de termos que podem ser selecionadas para o preenchimento do campo, reduzindo assim que sejam cadastrados mesmos dados, mas de diferentes formas. Sobre cada um dos dados extraídos foi realizado um levantamento possibilitando assim a descrição a seguir.

No que se refere à Formação, identificaram-se 377 formações diferentes. Este elevado número se deve à própria forma de preenchimento do campo na Plataforma Lattes. Devido a este ser um campo aberto permite diversas formas de escrita, como o uso ou não de acentos e especificações. Por exemplo, na área de

formação: Química, foram encontrados 23 diferentes formas de cadastramento na base, conforme Quadro 7:

QUADRO 7 - FORMAS DE CADASTRO DE UMA MESMA AREA NA BASE

- bacharelado em química tecnológica;
- bacharelado em química;
- bacharelado e licenciatura em química;
- bacharel em química;
- bacharel e licenciatura em química;
- licenciatura em ciências químicas;
- bacharel em química com atribuições tecnológicas;
- licenciatura em química;
- licenciatura em ciências habilitação em química;
- ciências química;
- química industrial;
- licenciatura e bacharelado em química;
- química;
- bacharelado em química com atribuições tecnológica;
- licenciatura plena em química;
- química - bacharelado, industrial, biotecnológico;
- química bacharelado com atribuições tecnológicas;
- bacharelado em química;
- licenciatura em química;
- bacharelado e licenciatura em química;
- bacharel em química;
- licenciatura e bacharelado em química;
- quimico laboratorista.

FONTE: Dados da pesquisa (A Autora, 2013)

Referente aos grupos de trabalhos, neste caso os departamentos, identificou-se que a quantidade total de diferentes departamentos cadastrados é 116 e que 349 pesquisadores não preencherem este dado. O número máximo de pesquisadores em um mesmo departamento é 31 pesquisadores, mas houve departamentos com apenas um único pesquisador cadastrado. No geral, a média de pesquisadores por departamento é 6,24, tendo como desvio padrão 6,52 e uma variância de 42,56 pesquisadores por departamento.

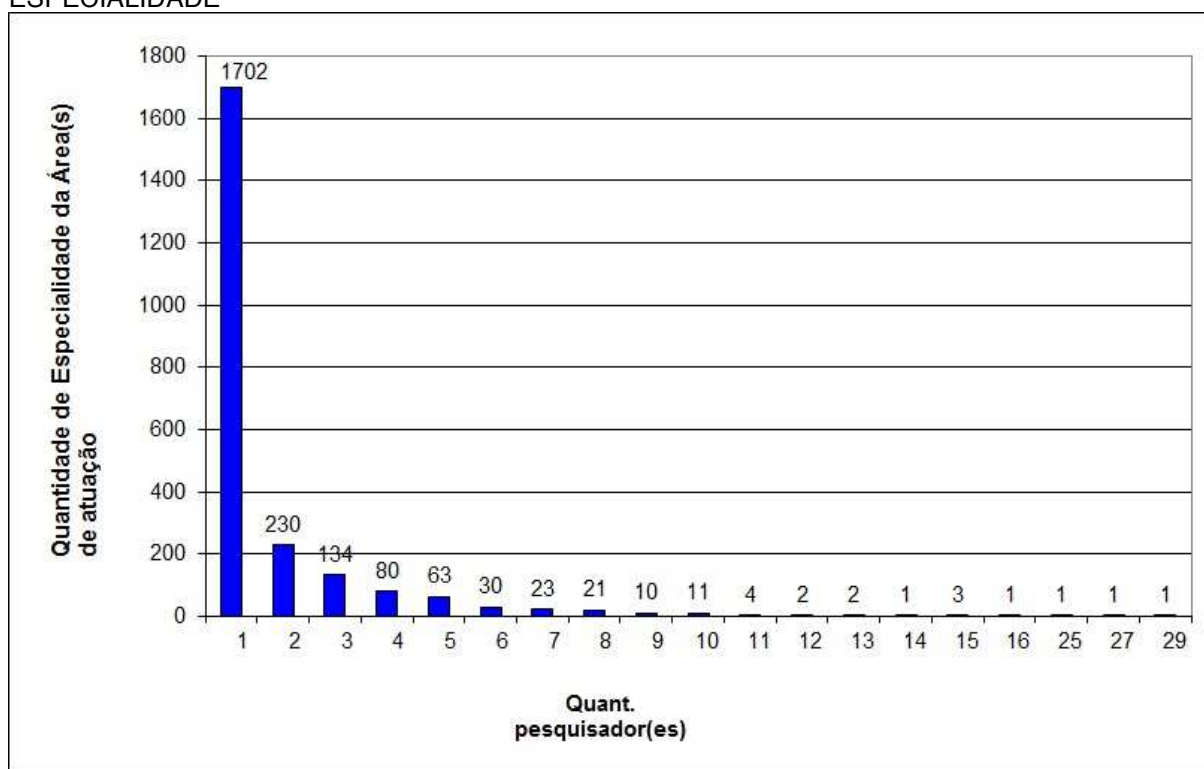
Sobre a formação, a quantidade máxima de graduações completas identificadas para um mesmo pesquisador foi três, sendo que 13 pesquisadores atendem a esta situação e 101 outros possuem duas graduações completas, este grupo com mais de uma formação representa 11% da base.

No total, 10% dos pesquisadores não informaram nenhuma graduação. A formação que mais possui pesquisadores na base conta com 55 deles. Há 267 graduações com apenas um pesquisador. A maior parte destas situações pode ser atribuída ao fato deste campo ser preenchido de forma livre. Com a obtenção destes

números identificou-se que a média de pesquisadores por formação é 3,06, com um desvio padrão igual a 6,2 e variância por formação igual a 38,44.

Com referência à Área de atuação, no total foram informadas 2.320 áreas de atuação diferentes considerando o campo Especialidade. A base Lattes permite que o usuário registre em cada área de atuação até quatro níveis: grande área, área, subárea do conhecimento e especialidade. Portanto, a Especialidade é o campo mais específico da área de atuação. O Gráfico 1 apresenta a relação entre pesquisadores em uma mesma área de atuação, considerando especificamente a máxima categorização permitida no Lattes, que é a Especialidade da área de atuação. No eixo vertical consta a quantidade de áreas de atuação, a especialidade desta, e no eixo horizontal a quantidade de pesquisadores que informaram aquela determinada especialidade. Por exemplo, a primeira coluna do Gráfico 1 aponta que existem 1.702 especialidades com apenas um único pesquisador cadastrado e o número máximo de atores em uma mesma especialidade é de 29 pesquisadores. O Gráfico 1 também permite identificar que existem apenas 18 especialidades com dois ou mais pesquisadores na mesma área.

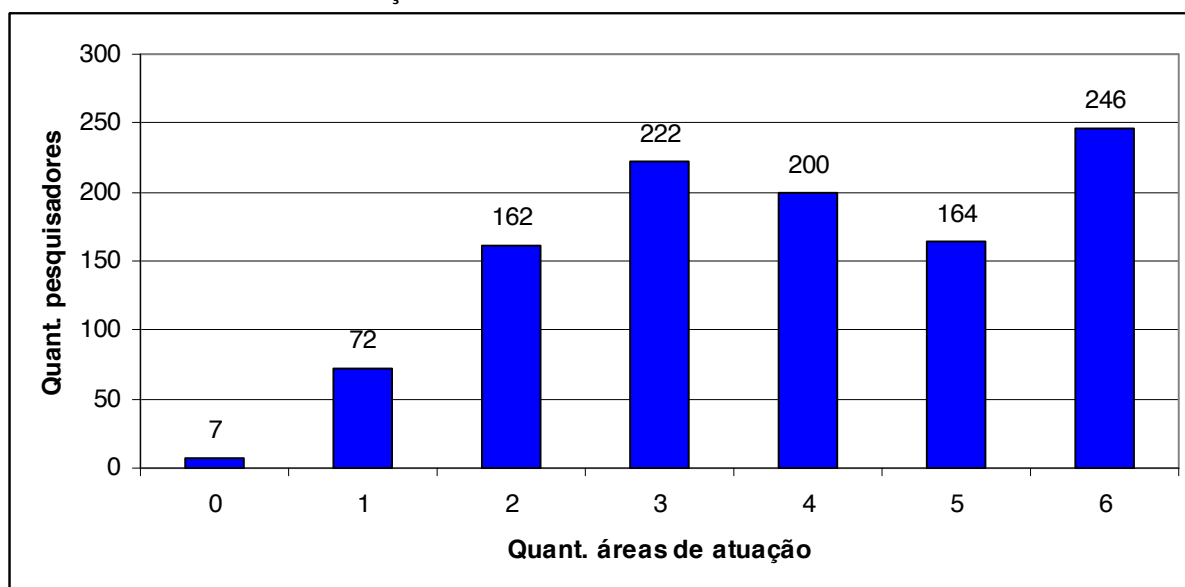
GRÁFICO 1 – PESQUISADORES EM UMA MESMA ÁREA DE ATUAÇÃO CONSIDERANDO A ESPECIALIDADE



FONTE: Dados da pesquisa (A Autora, 2012)

A Plataforma Lattes, até a data em que os dados foram coletados, vinte e dois de dezembro de 2011, permitia que cada pesquisador tivesse no máximo seis áreas de atuação cadastradas. Como este não é um campo obrigatório pode não ser preenchido, mas, na base, identificou-se que a maioria (22,9%) dos pesquisadores preencheu a quantidade máxima, ou seja, seis áreas de atuação. No Gráfico 2 é possível visualizar esta informação da quantidade de áreas de atuação que os pesquisadores cadastraram. No eixo vertical do gráfico consta o número de pesquisadores e no eixo horizontal a quantidade de áreas de atuação cadastradas. Observa-se, na primeira coluna que sete pesquisadores não cadastraram nenhuma área de atuação.

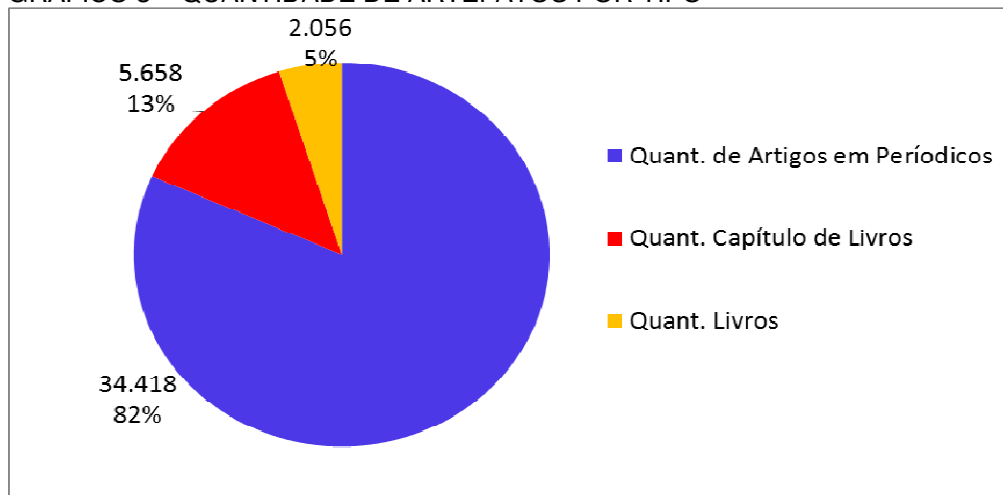
GRÁFICO 2 – ÁREAS DE ATUAÇÃO X PESQUISADORES



FONTE: Dados da pesquisa (A Autora, 2012)

A base tem cadastrados 39.142 artefatos que compreendem todos os artigos publicados em periódicos, capítulos de livros e livros. O total de cada tipo de artefato pode ser visualizado no Gráfico 3, onde identifica-se que Artigo é o tipo mais significativo nesta base, representando 82% dos artefatos cadastrados.

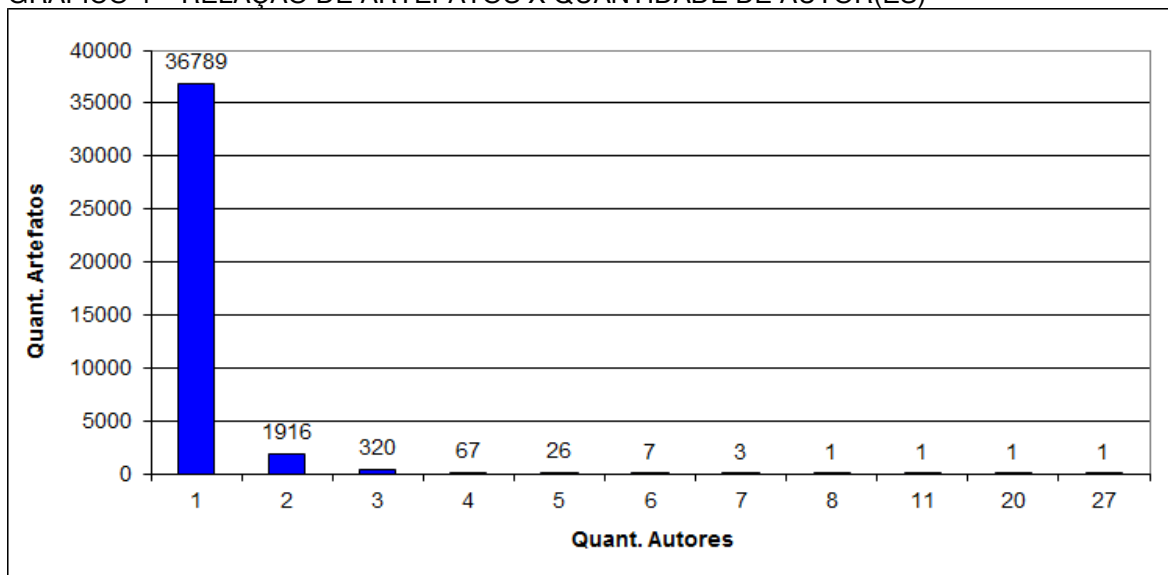
GRÁFICO 3 – QUANTIDADE DE ARTEFATOS POR TIPO



FONTE: Dados da pesquisa (A Autora, 2012)

Foram localizados 11 pesquisadores que não cadastraram nenhum artefato. O número máximo de artefatos de um mesmo ator foi de 835 artefatos, mas em média cada ator tem 39,64 artefatos, sendo que o desvio padrão é igual a 48,39 e a variância igual a 2342,42. No que se refere a produções de forma colaborativa, identificou-se que o número máximo de autores em um mesmo artefato foi de 27 autores, identificado em apenas uma publicação e que a grande maioria (87%) dos artefatos possui apenas um único autor. O Gráfico 4 apresenta esta relação de artefatos por autores, sendo que o eixo vertical mostra o número de artefatos e o eixo horizontal o número de autores. Por exemplo, na primeira coluna tem-se 36.789 artefatos cadastrados com um único autor.

GRÁFICO 4 – RELAÇÃO DE ARTEFATOS X QUANTIDADE DE AUTOR(ES)



FONTE: Dados da pesquisa (A Autora, 2012)

No que se refere às palavras-chave cadastradas, foram identificados 37.457 delas para os 39.132 artefatos cadastrados. Identificou-se também que 8.766 artefatos, representando 22% do total, não têm nenhuma palavra-chave associada. O número máximo de palavras-chave cadastradas para um mesmo artefato foi 18. Considerando-se o valor mínimo de uma palavra-chave, a média de palavras-chave por artefato é 2,73, com um desvio padrão igual a 0,85 e uma variância igual a 0,72.

Por meio da caracterização da base Lattes foi identificada a heterogeneidade da base Lattes, sendo que parte desta situação é atribuída à falta de padronização dos dados, pois muitos campos são de preenchimento aberto para o usuário realizar o cadastramento de seu currículo na Plataforma Lattes.

4.6 MINERAÇÃO DOS DADOS DA BASE LATTES

Com o propósito de identificar áreas de atuação/conhecimento relacionadas ao tema que está sendo pesquisado, a fim de auxiliar na busca por um conhecimento específico no mapa, técnicas de mineração de dados foram aplicadas para identificação destas áreas e também de outras relações não identificadas até o momento com a caracterização da base.

A tarefa de associação foi utilizada por possibilitar a identificação de associações não visíveis entre registros de dados que estão ou podem estar relacionados, desta forma pretende-se identificar as relações entre as áreas de atuação. Já a tarefa de classificação foi aplicada para descobrir o relacionamento entre o atributo meta que, refere-se à área de atuação e ao conjunto de atributos previsores, de modo que o processo de classificação utilize esse relacionamento para predizer a classe de um exemplo novo e não conhecido.

A seção 4.6.1 apresenta a fase de limpeza dos dados, que é realizada quando se identifica a necessidade de tratamento dos dados nos quais são identificados ruídos. Nas seções 4.6.2 a 4.6.4 são apresentados os resultados obtidos utilizando três algoritmos de mineração de dados.

4.6.1 Limpeza dos dados e pré-processamento

A mineração dos dados da base utilizada foi realizada visando apoiar o modelo proposto, pelas possíveis relações e associações descobertas com o uso desta técnica na base de dados a ser aplicada à prova de conceito. Por meio do resultado, é possível, por exemplo, desenvolver novas opções de buscas por conhecimentos relacionados por áreas que estejam associadas ao termo procurado, permitindo assim otimizar a localização de conhecimento relacionados e/ou associados.

No primeiro momento, foi definido um conjunto de dados alvo, inicialmente com 11 atributos, conforme descrição a seguir:

- “grupo_trabalho” – contém a informação do departamento;
- “formacao” – apresenta a primeira graduação concluída;
- “area_atuacao1”, “area_atuacao2”, “area_atuacao3”, “area_atuacao4”, “area_atuacao5”, “area_atuacao6” - áreas de atuação registradas, considerando o campo Especialidade;
- “artigos_3anos”, “cap_3anos”, “livros_3anos” - número total de artigos, capítulos de livros e livros com ano de publicação nos últimos 3 anos, tendo em vista que, as avaliações da Capes consideram os últimos 3 anos, e que os dados da base foram coletados no final do ano de 2011, consideraram-se os documentos publicados de 2009 a 2011. A categorização obedece ao formato:
 - “artigos_3anos” = zero; menor ou igual a 6 (≤ 6); maior que 6 (> 6);
 - “cap_3anos” = zero; um; maior ou igual a 2 (≥ 2);
 - “livros_3anos” = zero; um; maior ou igual a 2 (≥ 2).

O primeiro resultado obtido, executando o algoritmo de associação Apriori identificou que das 1.078 regras geradas, tendo como suporte mínimo de 0,1 (10%) e uma confiança mínima de 0,9 (90%) todas estavam relacionadas a áreas de atuação não preenchidas e/ou com nenhuma publicação registrada.

O algoritmo de classificação, C4.5 também obteve resultados não satisfatórios chegando a um percentual de erro das instâncias classificadas de 81%. A média de instâncias classificadas corretamente tendo como atributo meta as três

primeiras áreas de atuação foi 37,66% e conforme a matriz de confusão identificou-se que a média das três áreas classificadas 97,87%, portanto praticamente todas as instâncias classificadas corretamente, referia-se ao campo nulo. Esta situação é explicada por esta ser a categoria mais representativa na base.

O algoritmo de classificação, PRISM efetuou o teste de validação cruzada de 10 partições apenas quando os atributos meta eram: área de atuação 3 e área de atuação 6. Quando as demais áreas de atuação foram apontados como meta o algoritmo apenas apresentou as regras, sem conseguir realizar o teste de validação tendo o valor nulo como destaque.

Devido aos resultados preliminares apresentados não serem satisfatórios, foi identificada a necessidade de realizar a limpeza dos dados, conforme a seguir:

- Áreas de Atuação, buscando a generalização passou-se a considerar a área do conhecimento em detrimento à especialidade, como este atributo é de vocabulário controlado, diferentemente da especialidade, também se esperava que a incidência de termos diferentes, mas com o mesmo significado (sinonímia) e de campos nulos fosse reduzida.

- Departamento: como este é um campo de preenchimento aberto optou-se por padronizar os termos e adicionar a coluna setor, este se baseando no departamento informado, utilizando a relação de departamentos e setores da própria instituição. A padronização realizada está disponível no Apêndice A. Como o setor é uma generalização de departamento optou-se por excluir este campo.

- Formação: como este é um campo aberto, sem o uso de vocabulário controlado, optou-se por padronizar os termos e generalizá-los, para isto, foi considerado a nomenclatura principal da formação. A tabela com a padronização das formações está listada no Apêndice B.

- Demais atributos: os demais atributos referente a produtividade (número de artigos, capítulos de livros e livros total e dos últimos 3 anos) não foram alterados.

Para tratar o ruído oriundo dos campos não preenchidos e que, portanto, estavam como nulos, realizou-se o seguinte levantamento de campos nulos por atributo:

- Departamento = Total de 352 registros nulos;
- Formações = Total de 43 registros nulos;
- Área de Atuação 1 = Total de 8 registros nulos;
- Área de Atuação 2 = Total de 80 registros nulos;

- Área de Atuação 3 = Total de 241 registros nulos;
- Área de Atuação 4 = Total de 465 registros nulos;
- Área de Atuação 5 = Total de 663 registros nulos;
- Área de Atuação 6 = Total de 827 registros nulos.

Com base nestes números optou-se por excluir os 352 registros que estavam como nulos referente ao atributo departamento, passando a base para 721 registros. No que se refere às áreas de atuação, após a exclusão dos registros com departamento nulo ainda identificou-se a seguinte quantidade de registros:

- Área de Atuação 1 = 6 nulos;
- Área de Atuação 2 = 51 nulos;
- Área de atuação 3 = 172 nulos;
- Área de Atuação 4 = 324 nulos;
- Área de Atuação 5 = 465 nulos;
- Área de Atuação 6 = 563 nulos.

Devido ao elevado número de valores nulos principalmente nas três últimas áreas de atuação optou-se por retirar os atributos da área de atuação 4, 5 e 6, a base, portanto passou a contar com oito atributos. Setor, Formação, Área de Atuação 1, Área de Atuação 2, Área de Atuação 3, Quantidade de Artigos dos últimos 3 anos, Quantidade de Capítulos de Livros dos últimos 3 anos, Quantidade de Livros dos últimos 3 anos. As seções 4.6.2 a 4.6.4 apresentam os resultados após este tratamento dos dados, com a aplicação dos algoritmos de classificação e associação.

4.6.2 Resultado do algoritmo de associação: Apriori

Considerando o suporte mínimo de 0,1 (10%), e confiança maior que 0,9 (90%), o algoritmo encontrou oito regras, mesmo com a quantidade solicitada sendo superior (100 regras) e tendo em vista que, o algoritmo interrompe a execução quando a quantidade de regras é atingida ou o valor do suporte mínimo ou confiança

mínima é alcançada. As regras encontradas com o algoritmo estão apresentadas no Quadro 8.

QUADRO 8 - REGRAS DE ASSOCIAÇÃO ENCONTRADAS COM O ALGORITMO APRIORI

SE	E	ENTÃO	CONFIANÇA	OBSERVAÇÃO
Setor = Exatas	Quant. de capítulos = 0	Quant. de livros = 0	0,98	Das 82 ocorrências encontradas a regra se aplicou a 80.
Área de Atuação 3 = nulo	Quant. de capítulos = 0	Quant. de livros = 0	0,94	Das 109 ocorrências encontradas a regra se aplicou a 103.
Quant. de artigos >6	Quant. de capítulos = 0	Quant. de livros = 0	0,94	Das 139 ocorrências encontradas a regra se aplicou a 131.
Setor = Exatas	--	Quant. de livros = 0	0,94	Das 112 ocorrências encontradas a regra se aplicou a 105.
Setor = Biológicas	--	Quant. de livros = 0	0,93	Das 112 ocorrências encontradas a regra se aplicou a 104.
Setor = Tecnologia	--	Quant. de livros = 0	0,93	Das 84 ocorrências encontradas a regra se aplicou a 78.
Quant. de capítulos = 0	--	Quant. de livros = 0	0,92	Das 393 ocorrências encontradas a regra se aplicou a 360.
Quantidade de artigos <=6	Quant. de capítulos = 0	Quant. de livros = 0	0,91	Das 200 ocorrências encontradas a regra se aplicou a 181.

FONTE: A Autora (2013)

Por meio do Quadro 8 identifica-se que todas as regras encontradas com este algoritmo se referem à quantidade de livros publicados nos últimos três anos, apontando que dos dez setores existentes na base (Ciências Humanas, Letras e Artes; Ciências Jurídicas; Ciências da Saúde; Ciências Biológicas; Tecnologia; Ciências da Terra; Ciências Agrárias; Ciências Exatas; Ciências Sociais Aplicadas e Educação), dois tiveram uma produção menor de livros que o identificado nos demais setores, além de terem sido encontradas associações com a quantidade de artigos e de capítulos de livros. No entanto, nenhuma regra foi gerada referente às áreas de atuação.

4.6.3 Resultado do algoritmo de árvore de decisão: C4.5

Da aplicação do algoritmo C4.5 considerando como atributo meta a Área 1 e utilizando-se validação cruzada de 10 partições, foi induzida uma árvore com 910 folhas e tamanho 928.

O atributo área de atuação 1 foi escolhido como meta devido ao objetivo da mineração de dados contribuir para o mapa do conhecimento, onde espera-se que, por meio desta técnica possa-se encontrar relações entre as áreas de atuação. A área de atuação 1, que é a primeira área de atuação cadastrada pelo pesquisador, também foi a que apresentou maior quantidade de instâncias classificadas corretamente (70,73%) comparando com a área de atuação 2 (69,07%) e área de atuação 3 (55,61%).

A matriz de confusão que apresenta os acertos por classes está no Quadro 9, como a base contém 63 diferentes elementos relativos à área de atuação 1, a matriz de confusão gerada (tamanho 63 X 63) será apresentada parcialmente, apenas com os 5 elementos que obtiveram a maior precisão.

QUADRO 9 - MATRIZ PARCIAL DE CONFUSÃO DO CLASSIFICADOR OBTIDO PELO ALGORITMO C4.5

Classificado como →	Filosofia (0,933)	Psicologia (0,933)	Linguística (0,917)	Ciência Política (0,889)	Antropologia
Filosofia	14	0	0	0	0
Psicologia	1	14	0	0	0
Linguística	0	0	11	0	0
Ciência Política	0	0	0	8	0
Antropologia	0	0	0	0	8

FONTE: A Autora (2013)

Abaixo a sequência da relação dos atributos classificados incorretamente no Quadro 9, cujas áreas não constam na planilha:

- Filosofia = 2 instâncias classificadas 1 no setor de Educação e 1 em Zoologia;
- Psicologia = 4 instâncias classificadas como Educação;
- Linguística = 2 instâncias classificadas incorretamente, 1 Fisiologia e 1 como nulo;

- Ciência Política = 2 instâncias classificadas como Comunicação;
- Antropologia = 1 instância classificada como Medicina.

A estrutura da árvore gerada pelo algoritmo C4.5 para a base citada foi convertida para um conjunto de regras, no Quadro 10 é possível visualizar as principais regras de decisão encontradas (com maior número de elementos classificados).

QUADRO 10 - PRINCIPAIS REGRAS DE CLASSIFICAÇÃO GERADA PELO C4.5

SE	E	ENTÃO A ÁREA 1:
Setor = Ciências Sociais Aplicadas	Área 2 = Economia	Economia
Setor = Exatas	Área 2 = Química	Química
Setor = Exatas	Área 2 = Matemática	Matemática
Setor = Exatas	Área 2: Ciência da Computação	Ciência da Computação
Setor = Medicina	Formação: Medicina	Medicina

FONTE: A Autora (2013)

Buscando identificar especificamente as relações entre as áreas de atuação foram retirados da base os demais atributos e mantidos apenas os três que se referem à área de atuação, também foi mantido o atributo meta para a área de atuação 1, resultando em uma taxa de acerto de 70,87% (0,14% maior que quando considerado todas as instâncias) e reduzindo a árvore para 329 folhas e o tamanho para 334.

Especificamente buscando relações entre áreas diferentes, identificou-se 31 regras em que havia pelo menos uma instância que atendia a classificação realizada pelo algoritmo na base e que no mínimo uma das áreas não era a mesma, como exemplos a seguir:

- se a área de atuação 2: Oceanografia, então a área de atuação 1: Botânica;
- se área de atuação 2: Ciência e Tecnologia de Alimentos, então a área de atuação 1: Engenharia Química;

- se área de atuação 2: Planejamento Urbano e Regional, então a área de atuação 1: Geografia.

O quadro completo das regras encontradas, com as características descritas está no Apêndice C.

4.6.4 Resultado do algoritmo de regras de classificação: PRISM

Como abordado, o algoritmo PRISM tem por objetivo a indução de regras de classificação baseadas nas instâncias.

Considerando os mesmos parâmetros utilizados para o algoritmo C4.5, ou seja, atributo meta a área de atuação 1 e validação cruzada de 10 partições, o percentual de instâncias classificadas corretamente foi de 59,77%. O Quadro 11 apresenta a avaliação pela matriz de confusão apresentando os acertos por classes das classificações realizadas para os elementos com melhor índice de precisão.

QUADRO 11 - MATRIZ PARCIAL DE CONFUSÃO DO CLASSIFICADOR OBTIDO PELO ALGORITMO PRISM

Classificado como →	Filosofia (0,933)	Direito (0,917)	Física (0,875)	Educação (0,857)	Letras (0,833)
Filosofia	14	0	0	0	0
Direito	0	11	0	0	0
Física	0	0	14	0	0
Educação	0	0	0	24	0
Letras	0	0	0	0	10

FONTE: A Autora (2013)

Quanto às classificações errôneas destas instâncias apresentadas no Quadro 11, encontraram-se em:

- Filosofia = nenhuma instância classificada erroneamente;
- Direito = nenhuma instância classificada erroneamente;
- Física = uma instância classificada como engenharia elétrica;
- Educação = uma instância classificada como Direito, uma como Geociência, uma como Administração, uma como Química, uma como Enfermagem, uma como Educação Física (totalizando seis instâncias);

- Letras = uma instância classificada como Linguística e uma como nulo.

Diferentemente do C4.5 este algoritmo no Weka não apresenta a quantidade de instância da base classificada em cada uma das 354 regras encontradas pelo algoritmo para esta base, por este motivo serão apresentadas as primeiras regras geradas.

Ao manter na base somente as três instâncias referente às áreas de atuação, tendo como atributo meta a área de atuação 1, a quantidade de instâncias classificadas corretamente foi 56,17% e 309 regras, as dez primeiras regras, considerando ainda pelo menos uma das áreas diferentes são apresentadas no Quadro 12.

QUADRO 12 – PRIMEIRAS REGRAS DE CLASSIFICAÇÃO GERADA PELO PRISM

SE	E	ENTÃO ÁREA 1
Área 2 = Filosofia	Área 3 = Ecologia	Filosofia
Área 2 = Filosofia	Área 3 = Sociologia	Filosofia
Área 2 = Filosofia	Área 3 = História	Filosofia
Área 2 = Medicina	Área 3 = Saúde Coletiva	Medicina
Área 2 = Probabilidade e Estatística	Área 3 = Ciência da Computação	Medicina
Área 3 = Biologia Geral	Área 2 = Microbiologia	Medicina
Área 2 = Farmácia	Área 3 = Farmácia	Medicina
Área 2 = Educação	Área 3 = Nulo	Medicina
Área 3 = Nutrição	Área 2 = Psicologia	Fisiologia
Área 3 = Fisiologia	Área 2 = Bioquímica	Fisiologia

FONTE: A Autora (2013)

Por meio do Quadro 12 pode-se identificar que novas relações foram encontradas com a adoção do PRISM. Comparando com o resultado do C4.5, no entanto, o C4.5 apresentou ser o algoritmo mais adequado para esta base em virtude da maior taxa de classificações corretas (70,87% do C4.5 e 56,17% do PRISM) e comparando com o algoritmo Apriori que não encontrou nenhuma relação entre as áreas de atuação.

A seção 5 apresenta os resultados obtidos com a implementação do modelo por meio da prova de conceito desenvolvida utilizando os dados da Plataforma Lattes, bem como a estrutura de tabelas do MER proposto aplicando-o a esta base.

5 APLICAÇÃO DO MODELO

Esta seção apresenta os resultados obtidos com a implantação do modelo proposto aplicado aos dados dos currículos Lattes de professores/pesquisadores dos cursos *stricto sensu* da UFPR. A apresentação da base utilizada nesta prova de conceito, bem como os resultados obtidos com a mineração de dados e a caracterização da base de dados, especificamente dos componentes de entrada do modelo, foram descritos na seção 4. A presente seção tem como propósito apresentar e analisar a prova de conceito do mapa do conhecimento proposto a fim de analisar os pontos fortes e fracos do modelo.

5.1 PROCESSO DE CONSTRUÇÃO DA PROVA DE CONCEITO

A construção de uma prova de conceito, por meio da aplicação do modelo proposto na Universidade Federal do Paraná, permitiu a avaliação e a identificação dos pontos fortes e fracos no modelo.

A prova de conceito foi desenvolvida em duas fases. Na primeira fase, referente às etapas 1 e 2 do modelo apresentado no Quadro 5, utilizou-se a linguagem de programação Java em conjunto com a biblioteca Xstream para ler e mapear os dados originais disponíveis em formato de arquivo XML para banco de dados relacional. A biblioteca Xstream é capaz de ler dados em formato XML e transformá-los em objetos da linguagem Java. Após a extração, os dados foram inseridos no banco de dados PostgreSQL. A estrutura de tabelas criada no banco de dados respeita a modelagem apresentada na seção 4.4.

Para a fase de mapeamento das informações existentes para o modelo, utilizaram-se as seguintes regras:

Ator: o mapeamento dos dados do arquivo XML para o modelo proposto utiliza o atributo NUMERO-IDENTIFICADOR da marcação CURRICULO-VITAE associado ao campo identificador da tabela ator.

```
<CURRICULO-VITAE ... NUMERO-IDENTIFICADOR="XXXX">
...
</CURRICULO-VITAE>
```

Grupo de Trabalho: no mapeamento do Grupo de Trabalho do modelo proposto para os dados disponíveis nos arquivos XML, optou-se por associar o Grupo de Trabalho com o departamento informado pelo pesquisador/professor. Entretanto, verifica-se que não existe uma padronização no preenchimento deste dado (departamento), e observa-se na base a existência de currículos com o nome do departamento informado no atributo NOME-ORGÃO da marcação ENDERECO-PROFISSIONAL, mas em outros, surge o nome do departamento no atributo NOME-UNIDADE ou NOME-INSTITUICAO-EMPRESA.

```
<CURRICULO-VITAE ...>
  <DADOS-GERAIS ...>
    <ENDERECO ...>
      <ENDERECO-PROFISSIONAL ... NOME-INSTITUICAO-EMPRESA="XXX"
NOME-UNIDADE="YYY" NOME-ORGÃO="ZZZ"/>
    </ENDERECO>
  ...</DADOS-GERAIS>
</CURRICULO-VITAE>
```

Desta forma, definiu-se o procedimento a seguir para realização deste mapeamento:

Lê atributo NOME-ORGÃO

Se NOME-ORGÃO começa com texto “Dep”

Então

Grupo de Trabalho = NOME-ORGÃO

Finaliza

Se NOME-UNIDADE começa com texto “Dep”

Então

Grupo de Trabalho = NOME-UNIDADE

Finaliza

Se NOME-INSTITUICAO-EMPRESA começa com texto “Dep”

Então

Grupo de Trabalho = NOME-INSTITUICAO-EMPRESA

Finaliza

Formação: no mapeamento da Formação do modelo proposto para os dados disponíveis nos arquivos XML, optou-se por associar a Formação com as

graduações concluídas do pesquisador/professor. Nesta associação, o pesquisador/professor pode ter nenhuma, uma ou mais graduações cadastradas.

Área de conhecimento: no mapeamento da área de conhecimento do modelo proposto, optou-se por associar a área de conhecimento com as áreas de atuação informadas. O currículo Lattes limita o número de áreas de atuação cadastradas por pesquisador/professor até o máximo de seis. Desta forma, nesta associação, o pesquisador/professor pode ter nenhuma, uma ou mais áreas de atuação cadastradas.

Artefato: no mapeamento do Artefato do modelo proposto, optou-se por associar o Artefato com produção bibliográfica disponível no currículo Lattes. A produção bibliográfica está dividida em três tipos: artigos, capítulos de livros e livros. Também foi associada a data da produção bibliográfica com o artefato. Os autores da produção bibliográfica foram associados aos atores do modelo proposto. Como no modelo proposto um ator pode ter vários artefatos e um artefato pode ter vários autores, situação comum na produção acadêmica, o processo de mapeamento respeitou esta forma de associação, diferenciando produções bibliográficas com o mesmo título pelo atributo ano. As palavras-chave associadas à produção bibliográfica também foram associadas aos artefatos do modelo.

Após a fase de mapeamento, foi desenvolvida uma ferramenta de linha de comando em linguagem de programação Java para ler arquivos em formato XML da Plataforma Lattes. Para a leitura do conteúdo dos arquivos XML utilizou-se a biblioteca Xstream que permite o mapeamento do conteúdo XML para classes da linguagem Java. Este processo permitiu a extração somente dos dados desejados. Após a extração, utilizou-se a biblioteca JDBC (*Java Database Connection*), responsável por realizar a conexão com o banco de dados, da linguagem Java para inserir os dados no banco de dados relacional PostgreSQL. O PostgreSQL foi selecionado por ser um banco de dados compatível com a biblioteca Django, utilizada no desenvolvimento da interface web da prova de conceito.

Aplicando a estrutura de tabelas do MER, apresentado na Figura 20 e no Quadro 12, na base de currículos Lattes, as tabelas a seguir passaram a conter os seguintes dados:

- a) ator: identificador do pesquisador/professor;
- b) formacao: identificador da graduação e o nome do curso de graduação;

- c) grupotrabalho: identificador do departamento ao qual o ator faz parte e o nome deste departamento;
- d) artefato: identificador do título da produção científica e o tipo de artefato;
- e) tipoartefato: identificador referente ao tipo de artefato e nome do tipo de artefato. Por exemplo: artigo publicado em periódico, capítulo de livro, entre outros e o ano de sua publicação;
- f) palavrachave: identificador referente a palavra-chave e o atributo nome que se refere à própria palavra-chave.

Na segunda fase, responsável pelas etapas 3, 4 e 5 do modelo proposto apresentado no Quadro 5, após a fase de leitura e armazenamento dos dados da plataforma Lattes no banco de dados, foi desenvolvida a interface web da prova de conceito utilizando-se a linguagem de programação Python, a biblioteca Django e a biblioteca para criação de grafos JIT do inglês *JavaScript InfoVis Toolkit*. A linguagem Python foi utilizada pelo amplo suporte a bibliotecas e funcionalidades para desenvolvimento de ferramentas científicas. A biblioteca Django foi escolhida por permitir o desenvolvimento de sistemas em ambiente web, por possuir suporte para o banco de dados PostgreSQL e poder ser integrada com bibliotecas Javascript.

A opção por desenvolver em ambiente web foi baseada na possibilidade de permitir o acesso de diversos usuários sem a necessidade de instalar programas nos computadores. O acesso via ambiente exige somente o uso de um navegador web, programa geralmente instalado por padrão nos sistemas operacionais disponíveis atualmente.

O uso da biblioteca Django se deve as características como permitir a programação da prova de conceito de forma rápida, devido a diversas ferramentas que a biblioteca já possui, como a interface administrativa que permite a visualização, consulta, edição e exclusão de dados.

A interface web fez uso da biblioteca JIT para visualização dos grafos que representam os relacionamentos entre os elementos do modelo proposto.

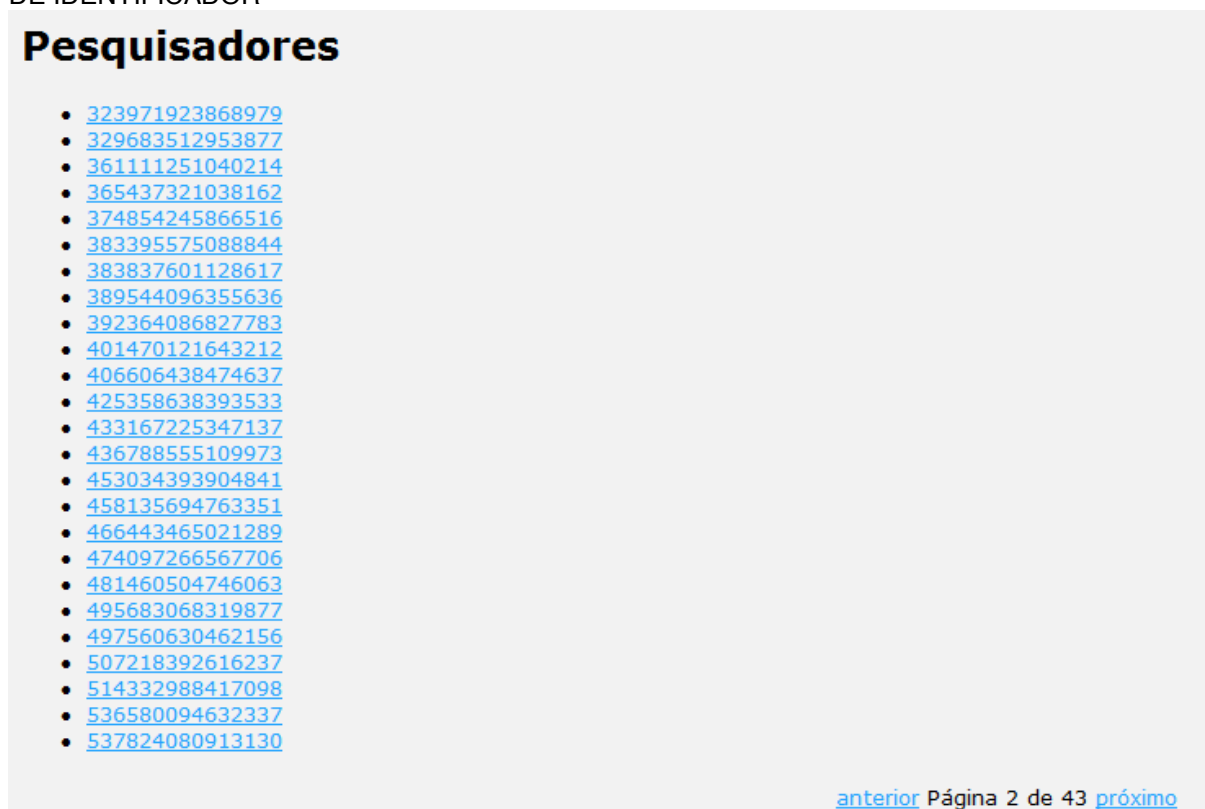
A seção 5.2 apresenta a interface da prova de conceito desenvolvida.

5.2 INTERFACE DA PROVA DE CONCEITO

A interface da prova de conceito foi desenvolvida para atender a etapa 3 (Quadro 5), disponibilização dos resultados do processo apresentado na seção 5.1 para acesso e consultas ao mapa. O usuário do sistema é capaz de acessar as informações disponíveis na base de dados e realizar consultas simples, como por pesquisador e área de atuação.

A Figura 21 apresenta a tela inicial do sistema onde os atores cadastrados são listados na forma de páginas amarelas. Visando proteger a identidade dos pesquisadores/professores optou-se por não apresentar o nome, expondo apenas o número identificador do arquivo. Também está disponível uma listagem de todas as áreas de atuação cadastradas no sistema. O usuário pode selecionar um ator, neste caso o identificador, para obter mais informações.

FIGURA 21 – TELA INICIAL DA PROVA DE CONCEITO COM LISTA DE ATORES POR NÚMERO DE IDENTIFICADOR



FONTE: A Autora (2013)

A Figura 22 mostra informações detalhadas sobre um ator selecionado apresentando as informações: grupo de trabalho associado (departamento), formações (graduação), áreas de conhecimento (áreas de atuação) e artefatos

produzidos (produção científica). Optou-se por passar as informações referente a: departamento, formação, área de atuação e títulos de artefatos para letras minúsculas, a fim de evitar diferenciações do sistema entre letras minúsculas e maiúsculas.

A produção científica é apresentada obedecendo a ordem: artigos, capítulos de livros e livros. Para cada um destes tipos de artefato utilizou-se o ano de publicação na ordem decrescente. Este critério foi adotado para que as publicações mais recentes sejam listadas por primeiro para o usuário.

FIGURA 22 – TELA COM INFORMAÇÕES DETALHADAS SOBRE UM DETERMINADO ATOR

Pesquisador																															
Identificador	514332988417098																														
Departamento	departamento de geologia																														
Graduações	curso de geologia																														
Áreas de Atuação	<table border="1"> <thead> <tr> <th>Grande Área</th><th>Área</th><th>Sub-área</th><th>Especialidade</th></tr> </thead> <tbody> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Geologia</td><td>geologia econômica;</td></tr> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Filosofia da Ciência</td><td>;</td></tr> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Geologia</td><td>geoquímica;</td></tr> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Geologia</td><td>geologia estrutural;</td></tr> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Ensino de Geologia</td><td>;</td></tr> <tr> <td>ciencias_exatas_e_da_terra</td><td>geociencias</td><td>Geologia</td><td>estratigrafia;</td></tr> </tbody> </table>			Grande Área	Área	Sub-área	Especialidade	ciencias_exatas_e_da_terra	geociencias	Geologia	geologia econômica;	ciencias_exatas_e_da_terra	geociencias	Filosofia da Ciência	;	ciencias_exatas_e_da_terra	geociencias	Geologia	geoquímica;	ciencias_exatas_e_da_terra	geociencias	Geologia	geologia estrutural;	ciencias_exatas_e_da_terra	geociencias	Ensino de Geologia	;	ciencias_exatas_e_da_terra	geociencias	Geologia	estratigrafia;
Grande Área	Área	Sub-área	Especialidade																												
ciencias_exatas_e_da_terra	geociencias	Geologia	geologia econômica;																												
ciencias_exatas_e_da_terra	geociencias	Filosofia da Ciência	;																												
ciencias_exatas_e_da_terra	geociencias	Geologia	geoquímica;																												
ciencias_exatas_e_da_terra	geociencias	Geologia	geologia estrutural;																												
ciencias_exatas_e_da_terra	geociencias	Ensino de Geologia	;																												
ciencias_exatas_e_da_terra	geociencias	Geologia	estratigrafia;																												
Artefatos	tipo	ano	palavras chave																												
	Artigo	2011	caracterização microestrutural do granito água comprida e formação de filonitos, antiforme setuva, paraná																												
	Artigo	2011	a experiência da oficina do mito à natureza: educar o olhar para as ciências da terra no festival de inverno de antonina (pr).																												
			o cristal, o ferro e o sal: recursos minerais do antigo quairá																												

FONTE: A Autora (2013)

A Figura 23 apresenta a forma de pesquisa por área de atuação a qual permite ao usuário procurar por todos os atores associados a uma determinada área. No escopo da prova de conceito as áreas de atuação foram mapeadas a partir das informações de grande área do conhecimento e área do conhecimento, disponíveis como campos, no currículo Lattes.

FIGURA 23 – TELA DE VISUALIZAÇÃO DO MAPA UTILIZANDO PESQUISA

<ul style="list-style-type: none"> 9991997365040337 9995803919265568 9996987729928765 9998364756956271 	
Pesquisas	
• Pesquisa por Área de Atuação	
CIENCIAS_AGRARIAS Agronomia nulo nulo	
CIENCIAS_AGRARIAS Agronomia nulo nulo	
CIENCIAS_AGRARIAS Agronomia Agroecologia nulo	
CIENCIAS_AGRARIAS Agronomia Agrometeorologia nulo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo nulo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo ciclagem de nutrientes	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo educação em solos	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo fauna do solo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo fertilidade do solo e adubação	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo gênese, morfologia e classificação dos solos	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo manejo e conservação do solo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo microbiologia e bioquímica do solo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo nutrição mineral de plantas	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo química do solo	
CIENCIAS_AGRARIAS Agronomia Ciência do Solo variabilidade espacial	
CIENCIAS_AGRARIAS Agronomia Desenvolvimento Rural Sustentável nulo	
CIENCIAS_AGRARIAS Agronomia Economia Rural nulo	
CIENCIAS_AGRARIAS Agronomia Extensão Rural desenvolvimento rural	
CIENCIAS_AGRARIAS Agronomia Extensão Rural ensino em solos	
CIENCIAS_AGRARIAS Agronomia Fitossanidade nulo	
CIENCIAS_AGRARIAS Agronomia Fitossanidade defesa fitossanitária	

FONTE: A Autora (2013)

O resultado da consulta pela área do conhecimento Ciência da Informação está apresentado na Figura 24. A partir deste resultado o usuário pode visualizar as demais especificações relacionadas a esta área do conhecimento (subáreas do conhecimento e especialidades) e também os pesquisadores que atuam nesta área, acessando assim as informações individuais de cada ator, permitindo desta forma a navegação por entre os componentes do modelo. Outro resultado desta consulta são as áreas relacionadas, se houver, obtidas por meio da mineração de dados.

FIGURA 24 – RESULTADO DA CONSULTA POR ÁREA DO CONHECIMENTO

Pesquisa por Área do Atuação			
Áreas do Conhecimento	<ul style="list-style-type: none"> ciencias_sociais_aplicadas ciencia_da_informacao ciencias_sociais_aplicadas ciencia_da_informacao Arquivologia organização de arquivos ciencias_sociais_aplicadas ciencia_da_informacao Biblioteconomia formação do profissional de Informação ciencias_sociais_aplicadas ciencia_da_informacao Biblioteconomia metodologia científica ciencias_sociais_aplicadas ciencia_da_informacao Biblioteconomia seleção e tratamento da informação ciencias_sociais_aplicadas ciencia_da_informacao Biblioteconomia técnicas de recuperação de informação ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação gerenciamento de acervos ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação metodologias para a gestão da informação ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação políticas de informação ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação redes colaborativas ciencias_sociais_aplicadas ciencia_da_informacao Gestão da Informação repositórios digitais ciencias_sociais_aplicadas ciencia_da_informacao GESTÃO DA INFORMAÇÃO ciencias_sociais_aplicadas ciencia_da_informacao GESTÃO DA INFORMAÇÃO linguagens de indexação ciencias_sociais_aplicadas ciencia_da_informacao GESTÃO DA INFORMAÇÃO monitoramento da informação ciencias_sociais_aplicadas ciencia_da_informacao Gestão do Conhecimento ciencias_sociais_aplicadas ciencia_da_informacao Inteligencia Competitiva ciencias_sociais_aplicadas ciencia_da_informacao Metodologia da Pesquisa ciencias_sociais_aplicadas ciencia_da_informacao Tecnologia da Informação ciencias_sociais_aplicadas ciencia_da_informacao Tecnologias da Informação e Comunicação gestão da informação ciencias_sociais_aplicadas ciencia_da_informacao Teoria da Informação teoria geral da informação ciencias_sociais_aplicadas ciencia_da_informacao Universidade e Educação Corporativa 		
Pesquisadores	<ul style="list-style-type: none"> 1216712773575109 1216712773575109 1216712773575109 1216712773575109 		

FONTE: A Autora (2013)

A Figura 25 apresenta outra forma de pesquisa, por palavras-chave, onde uma lista de todas as palavras-chave que contenham o termo utilizado na busca é exibida (Figura 26).

FIGURA 25 – PESQUISA POR PALAVRA CHAVE

Pesquisas

- Pesquisa por Área de Atuação
ciencias_agrarias || agronomia
Pesquisar
- Pesquisa por Palavra Chave
mineração
Pesquisar

Rede

FONTE: A Autora (2013)

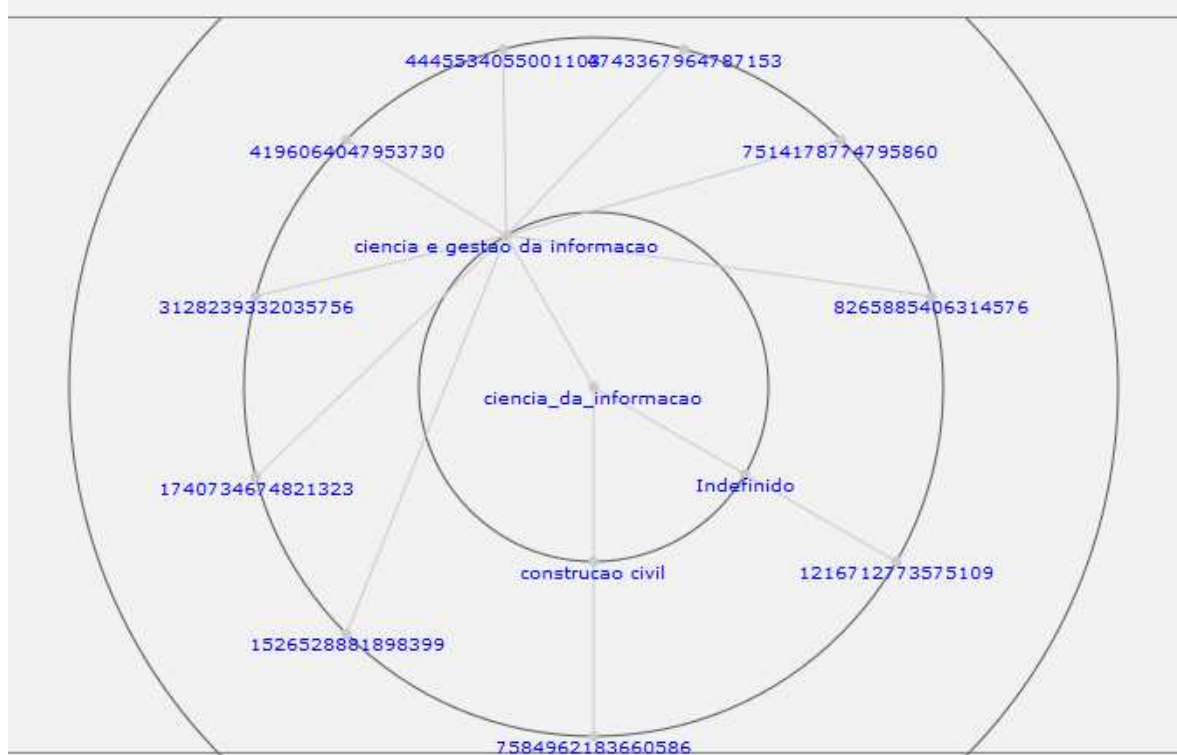
FIGURA 26 – RESULTADO DA PESQUISA POR PALAVRA CHAVE

Pesquisa por Palavra Chave	
Palavras Chave	<ul style="list-style-type: none"> • impactos da mineração • metais pesados, areas degradadas, mineração • mineração • mineração de dados • mineração de imagens
Voltar	

FONTE: A Autora (2013)

É possível, ainda, acessar uma quarta forma de visualização das informações disponíveis no banco de dados do mapa do conhecimento a disposição das áreas de atuação na forma de uma árvore, com centro a grande área do conhecimento e arestas ligando aos departamentos aos atores, como apresenta a Figura 27 cujo centro está a grande área: Ciência da Informação e as arestas ligam os departamentos e aos atores (representados pelo número identificador) que cadastraram em seu currículo esta grande área como área de atuação.

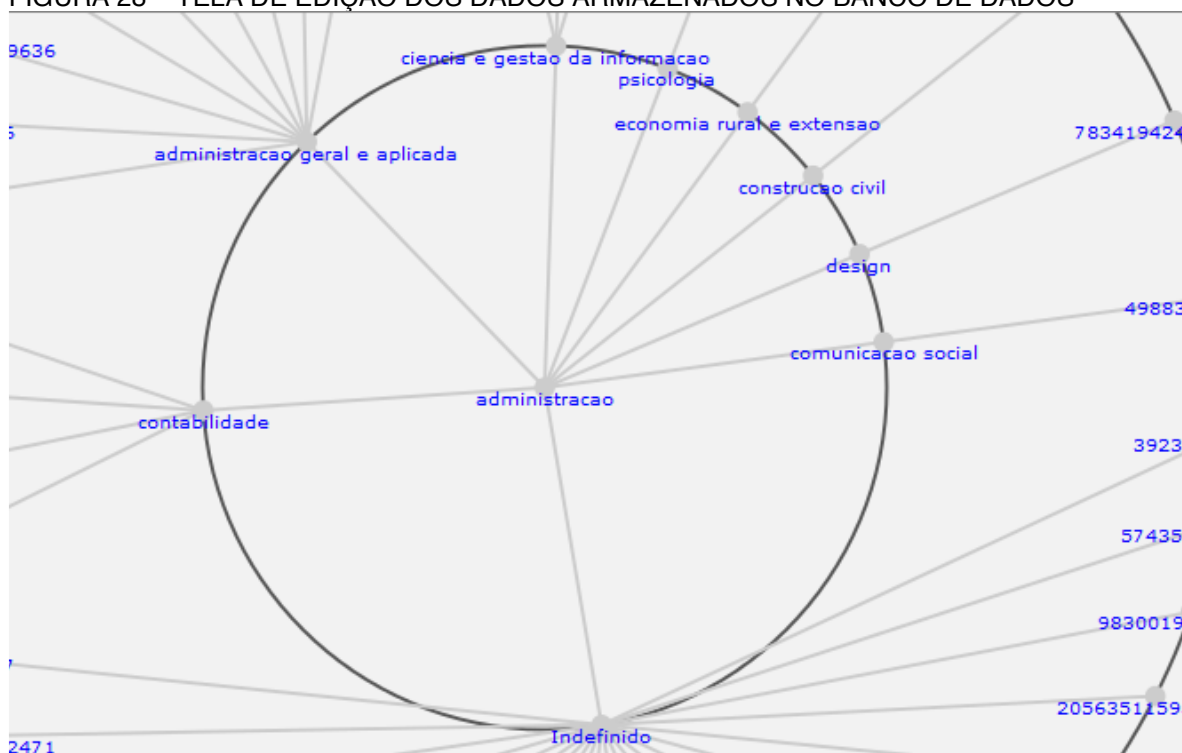
FIGURA 27 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS



FONTE: A Autora (2013)

A Figura 28 apresenta mais um exemplo desta quarta forma de visualização, tendo como centro da árvore a grande área: Administração.

FIGURA 28 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS



FONTE: A Autora (2013)

Estas árvores são geradas a partir de consultas realizada no banco de dados com uso da biblioteca JIT. É uma ferramenta dinâmica e interativa, que permite ao usuário selecionar um nó (UFPR, departamentos ou atores) para assumir o centro da árvore. Desta forma, encerra-se o experimento da etapa 4 do modelo proposto, relacionado a formas alternativas de visualização das informações.

Em relação a etapa 5 do modelo, disponibilização de sistema para edição e atualização das informações importadas para o banco, utiliza-se uma funcionalidade da biblioteca Django que permite a edição dos dados disponíveis no banco, conforme é apresentado na Figura 29.

FIGURA 29 – TELA DE EDIÇÃO DOS DADOS ARMAZENADOS NO BANCO DE DADOS

Django administration

Site administration

Auth	
Groups	+Add Change
Users	+Add Change

Componentes

Areas de atuacao	+Add Change
Artefatos	+Add Change
Atores	+Add Change
Formacoes	+Add Change
Grupos de trabalho	+Add Change
Palavras chave	+Add Change
Tipos de artefato	+Add Change

Sites

Sites	+Add Change
--------------	---

Recent Actions

My Actions

- [bacharelado em antropologia](#)
Formacao
- [antropologia](#)
Formacao
- [licenciatura em ciencias](#)
Formacao
- [meteorologia](#)
Formacao
- [tecnologia em gestão pública](#)
Formacao
- [licenciatura e formação de psicólogo](#)
Formacao
- [formação de psicólogos](#)
Formacao
- [administração de empresas](#)
Formacao
- [arqueologia](#)
Formacao
- [biociências](#)
Formacao

FONTE: A Autora (2013)

A funcionalidade apresentada na Figura 29 permite a edição das informações e também o acesso ao histórico de modificações, possibilitando assim o monitoramento do processo de atualização da base de dados.

5.3 ANÁLISE DOS RESULTADOS

Ao implementar o modelo proposto na base de dados Lattes foi possível identificar especificidades para aplicar a prova de conceito a esta base. Identificou-se que seria necessário o tratamento de algumas das situações encontradas, para que, ao disponibilizar a consulta simples ao usuário o mapa do conhecimento cumprisse seu objetivo facilitando a localização do conhecimento e a visualização das informações. Para isto, no mapeamento dos atributos para os componentes: ator, grupo de trabalho e artefato produzido, foi identificado que a falta de padronização comprometeria o resultado da aplicação da prova de conceito, desta forma, identificou-se a necessidade de tratamento dos dados de campos como: formação, áreas de atuação e departamento. Esta necessidade de tratamento na presente base pode ser atribuída aos campos de livre preenchimento, pois em se

tratando de uma base nacional, campos que poderiam ser de vocabulário controlado, auxiliando assim a padronização e reduzindo o uso de sinônimas, ou a ausência de dados por incerteza do seu correto preenchimento não o são, para possibilitar o seu uso em qualquer instituição de ensino e pesquisa em todo o território nacional e para atender aos mais diversos perfis de usuários. Em uma base mais homogênea, como em uma que seja utilizada especificamente em uma instituição, com orientações a seus usuários do correto preenchimento, o tratamento adotado aqui poderia ser dispensado.

Visando minimizar o impacto devido à ausência de padronização dos dados utilizados para os componentes de entrada do mapa do conhecimento, as soluções adotadas são analisadas a seguir.

Para o campo formação com o uso da padronização listada no Apêndice B, onde se utilizou uma nomenclatura única para cursos, desconsiderando especificações como tipo de habilitação do curso e diferentes formas de acentuação, identificou-se que para o mapa do conhecimento proposto a padronização adotada não ocasionou perda informacional e possibilitou a aplicação de algoritmos de mineração de dados resultando em um maior número de instâncias classificadas corretamente.

Para o grupo de trabalho também identificado como sem padronização, possibilitando com isso ser cadastrado em três diferentes campos da Plataforma Lattes, foi realizada a leitura de todos os campos até identificar a palavra “Dep” possibilitando que a informação fosse coletada do currículo, mas sem a necessidade de considerar os três campos simultaneamente, o que traria informação irrelevante para o mapa do conhecimento. Por outro lado, caso haja professor/pesquisador que não tenha cadastrado o departamento, utilizando a palavra “Dep”, esta informação não será recuperada passando o campo a ter o dado nulo. Entretanto, como foi identificado por meio de uma amostra que dos 352 registros nulos do campo grupo de trabalho os 36 primeiros registros identificados (representado 10% dos registros) não continham esta informação de departamento, portanto, também se avaliou que não houve perda informacional com a adoção deste critério.

Ainda com relação ao grupo de trabalho para sua padronização utilizou-se a lista de departamentos da UFPR, disponível no site desta instituição e listada no Apêndice A. Com este critério apenas em dois registros não foram identificados o

departamento correto, desta forma, avaliou-se que este critério de padronização foi adequado.

Com relação aos artefatos produzidos, em virtude de sua diferenciação ser pelos tipos (artigo publicado em periódico científico, capítulo de livro e livro), título e ano de publicação, caso haja mais de uma publicação do mesmo autor, do mesmo tipo, com o mesmo título e ano de publicação, este é considerado como único. O mesmo acontece caso outro ator cadastre um artefato com estes mesmos dados, este será identificado como um mesmo artefato, mas com mais de um autor e sabe-se que tal situação pode não ser a realidade, pela possibilidade deste se tratar de outra publicação. No entanto, mesmo assim por meio da aplicação nesta base este critério foi considerado satisfatório, pois ao realizar a caracterização, conforme apresentado no Gráfico 4 da seção 4.5, apenas quatro artefatos contaram com oito ou mais autores. No entanto, com este critério também foi possível identificar em se tratando de uma mesma publicação, esta deve ter sido cadastrada diferentemente por seus autores, considerando que 94% dos artefatos foram cadastrados por um único autor, o que para instituições acadêmicas em que a colaboração entre pesquisadores é constante, entende-se que uma mesma publicação pode ter sido cadastrada de forma diferente entre os autores explicando este elevado número de publicações com apenas um autor.

No que se refere à visualização do mapa, caso o número de atores seja elevado, ou seja, que dificulte a localização de um determinado ator, recomenda-se a apresentação da lista de atores na forma de páginas amarelas, em diferentes páginas, como por exemplo, por ordem alfabética apresentando o resultado apenas da letra selecionada, a fim de facilitar a localização dos atores pelos usuários.

A seção 6 apresenta as considerações finais realizadas sobre a pesquisa, o atendimento aos objetivos propostos e sugestões para trabalhos futuros.

6 CONSIDERAÇÕES FINAIS

Por meio da prova de conceito foi possível validar o modelo de mapa de conhecimento proposto, confirmando que este é aplicável e que pode contribuir para o compartilhamento do conhecimento na instituição, por apresentar uma forma de compartilhar o conhecimento explícito existente, das pessoas que atuam na instituição e identificar os pontos positivos e lacunas. Das vantagens provenientes da aplicação do modelo, destacam-se as características apresentadas a seguir.

Primeiramente, a criação de uma base de dados para o mapa do conhecimento. O modelo de entidade-relacionamento do banco de dados proposto permite a criação e alimentação de uma base de dados adequada para construção do mapa do conhecimento. O modelo proposto respeita as formas normais, evitando a repetição de dados iguais e permitindo que a alteração de dados, como o nome de um grupo de trabalho, reflita imediatamente sobre os registros relacionados.

Facilidade de navegação entre as informações cadastradas na base de dados, por meio de uma interface em ambiente web. A interface desenvolvida para prova de conceito permite que o usuário navegue entre os registros de atores e grupos de trabalho. A partir de um registro de ator, o usuário pode visualizar o grupo de trabalho ao qual o ator está associado, formação e os artefatos produzidos. A partir de um grupo de trabalho, o usuário pode visualizar os atores associados e os artefatos produzidos pelo grupo. Desta forma, todas as telas estão interligadas permitindo ao usuário navegar entre as informações. A interface também permite aos usuários alterar as informações cadastradas a fim de corrigir ou completar informações ausentes.

Uma terceira característica é a possibilidade de filtrar as informações e visualizá-las com diferentes técnicas. A interface apresentada na prova de conceito permite que o usuário filtre os registros visíveis por meio da inclusão de restrições nas consultas realizadas ao banco de dados. O sistema fornece a possibilidade de filtrar os atores por área de interesse. A interface também permite a visualização da associação das áreas do conhecimento com os grupos de trabalho e atores por meio de um grafo dinâmico, onde o usuário pode interagir com o grafo para definir qual elemento assume a posição central.

Por fim, referente aos grupos de trabalho identificam-se vantagens na possibilidade de inferir diferentes formas de relacionamento entre os grupos de trabalhos, devido à visualização do relacionamento entre as áreas de conhecimento/atuação identificadas pelo uso das técnicas de mineração de dados a fim de descobrir relacionamentos não visíveis em consultas simples.

Entre as dificuldades provenientes da aplicação do modelo proposto, destaca-se primeiramente a dificuldade para tratar dados derivados do preenchimento incompleto de currículos na Plataforma Lattes. Diversos registros analisados apresentam ausência de informações importantes para o modelo proposto como nome do departamento ao qual o pesquisador está associado (é mapeado como um grupo de trabalho do ator) ou, quando fornecido, não segue a nomenclatura oficial da universidade. Exemplos deste tipo de problema no cadastro são: informar apenas o endereço e não também o departamento, informar Lactec como departamento da UFPR (fisicamente o Lactec está dentro do campus da UFPR, mas não é um departamento), informar Pró-Reitoria (confundindo cargo assumido com departamento onde está lotado).

Outra dificuldade encontrada com o modelo refere-se às sinonímias, devido a todos os dados provenientes de campos de preenchimento aberto da base Lattes terem grande número de palavras diferentes, mas com o mesmo significado, em especial nos campos departamento, formação e palavras-chave. O modelo proposto não trata estas situações.

A seção 6.1 apresenta a síntese do atendimento aos objetivos propostos e sugestões para trabalhos futuros.

6.1 ATENDIMENTO AOS OBJETIVOS PROPOSTOS

Esta dissertação teve por objetivo geral desenvolver um modelo para a construção de mapa do conhecimento para instituições acadêmicas.

Por meio da revisão da literatura foram analisadas as técnicas de mapa do conhecimento, sendo identificada a definição de Jafari *et al.* (2009) como a mais adequada para atender ao objetivo e para alcançá-lo, foi necessário definir três objetivos específicos conforme apresentado na sequência.

- a) construir uma base de dados para armazenamento dos dados referente ao conhecimento de professores e pesquisadores e suas respectivas produções científicas.

A fim de armazenar os dados obtidos da fonte de dados original que seriam necessários para compor o mapa do conhecimento, foi definida a estrutura de tabelas e relacionamentos existentes entre atores, grupos de trabalho, formações, áreas de atuação, palavras-chave e dados auxiliares.

- b) identificar relacionamentos e associações nesta base de dados.

Visando a disponibilização dos resultados para os usuários do mapa, não apenas dos termos pesquisados, mas também de termos associados, técnicas de mineração de dados foram aplicadas para identificar associações e relações das áreas de conhecimento que eram desconhecidas até então, possibilitando assim retornar resultados de pesquisa das áreas de atuação não apenas da área pesquisada, mas também as áreas que estejam relacionadas.

- c) validar o método proposto, por meio da implementação de uma ferramenta para prova de conceito.

A prova de conceito desenvolvida permitiu além de validar o método identificar as vantagens, lacunas existentes e confirmar a viabilidade de implantação do modelo, atendendo assim ao propósito de facilitar a visualização da informação contribuindo para o compartilhamento de conhecimento entre professores e pesquisadores detentores de conhecimento, a fim de que o mesmo possa passar do individual ao coletivo.

Ainda neste enfoque, abrem-se leques de oportunidades para trabalhos futuros na perspectiva do aprimoramento do mapa do conhecimento, sugeridas a seguir.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

A partir das limitações expostas no presente trabalho no contexto específico da aplicação do mapa do conhecimento para instituições acadêmicas, sugere-se:

- recuperação de outros campos da base Lattes, como o número do DOI (*Digital Object Identifier*, padrão para identificação de

documentos na web) ou ISSN (*International Standard Serial Number*, identificador de publicações seriadas de aceitação internacional), que no ano de 2012 passou a ser validado a cada inclusão de artigo, possibilitando identificar estratos indicativos da qualidade (nível do Qualis, utilizado pela Capes) da produção;

- outras formas de pesquisa que podem ser implementadas, considerando a necessidade em outros ambientes ou atendendo a especificidades dos programas de mestrado e doutorado;
- aplicação do método proposto em outras bases de dados, a fim de realizar a comparação dos resultados da aplicação em diferentes ambientes;
- aplicação de outros métodos de mineração de dados que podem ser utilizados, como algoritmos de clusterização visando descobrir novos grupos entre os departamentos e pesquisadores;
- estudos de interface, a fim de identificar novas formas de apresentação para o modelo proposto.

Com estas sugestões para outros trabalhos que deem continuidade a este, encerra-se a presente pesquisa reforçando a viabilidade do método proposto e de sua contribuição para a gestão do conhecimento, por possibilitar uma forma de compartilhar o conhecimento explícito entre as pessoas atuantes na instituição em que o mapa do conhecimento é aplicado. Além de organizar e agrupar as informações em um único local referente a produção da instituição, possibilitando também a gestão de recursos.

REFERÊNCIAS

ANDREWS, K.; HEIDEGGER, H. Information slices: visualization and exploring large hierarchies using cascading, semi-circular discs. **Proceedings IEEE Symposium on Information Visualization** (InfoVis'98), out. 1998.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. **Proc. of the 20th Int'l Conference on Very Large Databases**. Santiago, Chile, 1994.

AGRAWAL, R.; MANNILA, H.; SRIKANT, R.; TOIVONEN, H. ;VERKANO. Fast Discovery of Association Rules. In: Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), **Advances in knowledge discovery and data mining**, Cambridge, Cap. 12, The MIT Press, 1996.

BARBOSA, R. Gestão da informação e do conhecimento: origens, polêmicas e perspectivas. **Informação & Informação**, Londrina, v. 13, p. 1-25, 2008.

BERRY, M; LINOFF, G. **Mastering data mining**: the art and science of customer relationship management. Nova Jersey: John Wiley & Sons, 1997.

CARD, S.K.; MACKINLAY, J.D.; SHNEIDERMAN, B. Information visualization. In: Card, S.K.; Mackinlay, J.D.; Shneiderman, B. (eds.). **Readings in information visualization - using visualization to think**. San Francisco: Morgan Kaufmann Publ., p. 1-34, 1999.

CAVALCANTI, M. B. C., GOMES, E. P. B., PEREIRA, A. **Gestão de empresas na sociedade do conhecimento**. Rio de Janeiro: Campus, 2001.

CENDROWSKA, J. PRISM: an algorithm for inducing modular rules. **International Journal of Man-Machine Studies**. p. 349-370, 1987

CFEngine AS. **Implementing knowledge management**: a CFEngine special topics handbook. 2009. Disponível em: <<https://cfengine.com/manuals/st-knowledge>>. Acesso em: 29 maio 2012.

CHI, E. H. A taxonomy of visualization techniques using the data state reference model. In: IEEE SYMPOSIUM ON INFORMATION VISUALIZATION, INFOVIS, 2000. **Proceedings...** [S.I.]: IEEE, 2000, p. 69-75.

CHOO, C.W. Information management for the intelligent organization: roles and implications for the information professions. **Digital Libraries Conference**, 1995. Singapore: National Computer Board of Singapore. p. 81-99. Disponível em: <<http://choo.fis.utoronto.ca/fis/respub/dlc95.html>>. Acesso em: 25 abr. 2012.

_____. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: SENAC, 2006.

CHUNG, F.; LU, L. **Complex graphs and networks**. Rhode Island: CBMS: 2006. 264p. Disponível em: <<http://www.math.ucsd.edu/~fan/complex/>>. Acesso em: 15 dez. 2012.

DJANGO. Disponível em: <<https://www.djangoproject.com/>>. Acesso em: 04 jan. 2013.

DAVENPORT, T. H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. 4. ed. São Paulo: Futura, 1998. 316p.

_____. **Reengenharia de processos**. Rio de Janeiro: Campus, 1994.

DAVENPORT, T. H., JARVENPAA, S. L., BEERS, M. C. Improving Knowledge Work Processes. **Sloan Management Review**, v. 37, p. 53-65, 1996.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Elsevier, 2003.

DIGIAMPIETRI, L.A. *et al.* Minerando e caracterizando dados de currículos lattes. In: **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**. 2012.

EBENER, S. *et al.* Knowledge mapping as a technique to support knowledge translation. **Bulletin of the World Health Organization**, p. 636-642, 2006.

FAYYAD, U.; SHAPIRO, P. G.; SMYTH, P. From data mining to knowledge discovery: an overview. In: Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), **Advances in knowledge discovery and data mining**, The MIT Press, p. 1-36, Cambridge, 1996.

_____. From data mining to knowledge discovery in databases. **American Association for Artificial Intelligence**, p. 37–54, 1996 b. Disponível em: <<http://www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf>>. Acesso em: 20 maio 2012.

FERREIRA, P. S. A. **Mapas do conhecimento**: em busca de um instrumento de aplicação nas organizações para operacionalização da gestão do conhecimento. [S.l.], 2007, 154 p. Dissertação (Curso de Mestrado em Gestão Empresarial). Fundação Getulio Vargas.

FRAWLEY, W.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G.; Frawley, W. (eds.), **Knowledge discovery in databases**, Cambridge: The MIT Press, p. 1-27, 1991.

FREITAS, C. M. D. S. *et al.* Introdução à visualização de informações. **Revista de Informática Teórica e Aplicada**. Porto Alegre, v. 8, n. 2, p. 143-158, 2001.

FOWLER, M. **UML essencial**: um breve guia para a linguagem-padrão de modelagem de objetos. Porto Alegre: Bookman, 2005.

GERSHON, N. **From perception to visualization, in scientific visualization**. Advances and Challenges, Ed: L. Rosenblum, R.A. Earnshaw, J. Encarnacao, H. Hagen, A. Kaufman, S. Klimenko, G. Nielson, F. Post, D. Thalmann, Academic Press. 1994.

GROTTO, D. O compartilhamento do conhecimento nas organizações. In: ANGELONI, M. T. **Organizações do conhecimento**: infra-estrutura, pessoas e tecnologias, 2003.

GUIMARÃES, C. C. **Fundamentos de banco de dados**: modelagem, projeto e linguagem SQL. Campinas: Unicamp, 2008.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, P. R.; WITTEN, I.H. **The weka data mining software: an update**. Disponível em: <<http://www.cs.waikato.ac.nz/~ml/weka/>>. Acesso em: 15 jan. 2013.

HAN, J; KAMBER, M. **Data mining**: concepts and techniques. San Francisco, 2 ed. 2006.

HANSEN, M; NOHRIA, N. TIERNY, T. What's your strategy for managing knowledge? **Harvard Business Review**, p. 106-116, mar./abr., 1999.

HOWARD, A. R. Knowledge maps. **Management Science**, v. 35. n. 8, ago. 1989.

HERMAN, I.; MELANÇON, G.; MARSHALL, M, S. Graph visualization and navigation in information visualization: a survey. **IEEE Transactions on visualization and computer graphics**, v. 6. 2000. p. 1-21.

JAFARI, M; *et al.* A Framework for the selection of knowledge mapping techniques. **Journal of Knowledge Management Practice**, v. 10, n. 1, 2009. Disponível em: <<http://www.tlinc.com/articl180.htm>>. Acesso em: 31 maio 2012.

JIT - JavaScript InfoVis Toolkit. Disponível em: < <http://philogb.github.com/infovis/>>. Acesso em 10 jan. 2013.

JOHNSON, B.; SHNEIDERMAN, B. TreeMaps: a space - filling approach to the visualization of hierarchical information structures. **Proceedings of IEEE Visualization**. San Diego, p. 284 - 291. 1991.

KEIM, D. A. Visual database exploration techniques. In: **International conference in knowledge discovery and data mining**, KDD, 1997. Disponível em:<<http://www.dbs.informatik.uni-muenchen.de/~daniel/KDD97.pdf>>. Acesso em: 12 ago. 2012.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence**. Morgan Kaufmann, San Mateo, California. p. 1137-1143, 1995.

LE-KHAC, N.-A.; AOUAD, L. M.; KECHADI, M.-T. An efficient support management tool for distributed data mining environments. In: **IEEE**, 2007.

_____. Knowledge map: Toward a new approach supporting the knowledge management in distributed data mining. In: **Third international conference on autonomic and autonomous systems (ICAS'07)**, 2007.

LIANG, Z.; XUEMING, L. An ontology reasoning architecture for data mining knowledge management. **Wuhan University Journal of Natural Sciences**, v. 13, p. 396–400, 2008.

MALDONADO, M. U. **Análise do impacto das políticas de criação e transferência de conhecimento em processos intensivos em conhecimento**: um modelo de dinâmica de sistemas. Florianópolis, 2008. 138f. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento). Universidade Federal de Santa Catarina.

MARCONI, M. de A.; LAKATOS, E. M. **Metodologia científica**. 5. ed. São Paulo: Atlas, 2007. 312 p.

MENA-CHALCO, J.P.; CESAR-JR, R.M. scriptLattes: An open-source knowledge extraction system from the lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009.

_____. Exemplos de relatórios do scriptLattes. 2012. Disponível em: <http://www.vision.ime.usp.br/creativision/publications_dcc/>. Acesso em: 17 jul. 2012.

MENEZES, V. S. A. **Análise de redes sociais científicas**. Rio de Janeiro, 2012. 222 p. Tese (Programa de Engenharia de Sistemas e Computação). Universidade Federal do Rio de Janeiro.

MICROSOFT. **O que é o Excel?** Disponível em: <<http://office.microsoft.com/pt-br/novice/o-que-e-o-excel-HA010265948.aspx>>. Acesso em: 26 jan. 2013.

MUGNAINI, R.; DIGIAMPIETRI, L. A.; OLIVEIRA, L. C. de; FERREIRA, S. M. S. P. Fontes de informação institucionais para normalização automática de nomes de autores: proposta de um método automático. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 3., 2012, Gramado. Anais... Gramado: UFRGS, 2012.

NONAKA, I; TAKEUCHI, H. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. Campus: Rio de Janeiro, 1997.

NORTH. C.A. **Taxonomy of information visualization user interfaces**. 1998. Disponível em: <<http://www.lirmm.fr/~infoviz/ASEval/References/shneiderman.php>>. Acesso em: 01 ago. 2012.

OLIVE - On-line library of information visualization environments. 1999. Disponível em: <<http://otal.umd.edu/Olive/>> Acesso em: 01 ago. 2012.

ORACLE. Disponível em:

<<http://www.oracle.com/technetwork/java/javase/overview/javahistory-index-198355.html>>. Acesso em: 10 jan. 2013.

PIROLA, V. S. **A proposição de uma ferramenta de apoio ao mapeamento do conhecimento em uma organização**. Rio de Janeiro, 2002, 130 p. Dissertação (Mestrado em Ciências em Engenharia de Produção). Universidade Federal do Rio de Janeiro.

PLATAFORMA Lattes. Disponível em: <<http://lattes.cnpq.br/>> Acesso em: 03 jan. 2013.

POSTGRESQL. **About**. Disponível em: <<http://www.postgresql.org/about/>> Acesso em: 16 jan. 2013.

PREGIBON, D. Data mining. In: HANSEN, M; PERUGGIA, M. (eds.), Statistical Computing and Graphics. American Statistical Association. v. 7, n. 3, Alexandria, december, 1996. Disponível em: <<http://stat-computing.org/newsletter/issues/scgn-07-3.pdf>>. Acesso em: 06 jun. 2012.

PROBST, G.; RAUB, S.; ROMHARDT, K. **Gestão do conhecimento**: os elementos construtivos do sucesso. Porto Alegre: Bookman, 2002.

PROPLAN – Pró-Reitoria de Planejamento, Orçamento e Finanças. **Gráficos evolutivos UFPR 1997-2011**. Disponível em: <<http://www.proplan.ufpr.br/home/CPI/arquivos/Gr%C3%A1ficos%20evolutivos%20UFPR%20abril%202012%20final.pdf>>. Acesso em: 01 out. 2012.

PRPPG/UFPR – Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Paraná. **Programas de mestrado e doutorado**. Disponível em: <http://www.prppg.ufpr.br/programas_pgmd>. Acesso em: 17 mar. 2012a.

_____. **Histórico da pós-graduação**. Disponível em: <http://www.prppg.ufpr.br/historico_pg>. Acesso em: 17 mar. 2012b.

PYTHON. Disponível em: < <http://www.python.org/>>. Acesso em: 28 dez. 2012.

QUINLAN, R. Induction of Decision Trees. In: SHAVLIK, Jude (ed.); DIETTERICH, Thomas (ed.). **Readings in Machine Learning**, San Mateo, CA: Morgan Kaufmann Publishers, p. 81-106, 1990.

REINGOLD; E. M.; TILFORD, J. S. Tidier Drawing of Trees, **IEEE Transactions on Software Engineering**, n. 2, p. 223–228, 1981.

RHYNE, T. M. Does the Difference between Information and Scientific Visualization Really Matter, **IEEE Computer Graphics and Applications**, maio/jun., p. 6-8, 2003.

ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. K. Cone trees: animated 3D visualizations of hierarchical information. In Robertson, S. P., Olson, G. M., and Olson, J. S. (eds.), **Proc. ACM Conf. Human Factors in Computing Systems**, CHI, p. 189–194. ACM Press. 1991.

SANTOS, M. F.; AZEVEDO S. C. **Data mining: descoberta de conhecimento em bases de dados**. Lisboa: FCA, 2005.

SASIETA, M. H. A. **Um modelo para visualização do conhecimento baseado em imagens semânticas**. Florianópolis, 2011, 207 p. Tese (Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento). Universidade Federal de Santa Catarina.

SAITO, D, S. **Investigação de técnicas de visualização para representação de autômatos finitos com saída**. Porto Alegre, 2003, 82 p. Dissertação (Programa de Pós-Graduação em Computação). Universidade Federal do Rio Grande do Sul.

SATO, K. A. S. **Criação e compartilhamento de conhecimento: o caso do projeto perfis profissionais para o futuro da indústria**. Curitiba, 2010, 185 p. Dissertação (Mestrado em Ciência, Gestão e Tecnologia da Informação). Universidade Federal do Paraná.

SCHLESINGER, C. C. B; REIS. D. R. dos; SILVA, H. de F. N. *et al.* **Gestão do conhecimento na administração pública**. 1. ed. Curitiba: Instituto Municipal de Administração Pública, 2008.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistemas de banco de dados**. 5. ed. Rio de Janeiro: Elsevier, 2006.

SILVA, G. **Mapeamento de conhecimento através das redes: visualizando o intangível**. TerraForum. Disponível em:
<<http://www.terraforum.com.br/biblioteca/Documents/Mapeamento%20de%20conhecimento%20atrav%C3%AAs%20das%20redes%20visualizando%20o%20intang%C3%ADvel.pdf>>. Acesso em: 06 set. 2012.

SILVEIRA, R. de F. **Mineração de dados aplicada à definição de índices em sistemas de raciocínio baseado em casos**. UFRGS, 2003.

SPENCE, R.; APPERLEY, M.D. Data base navigation: an office environment for the professional. **Behaviour and Information Technology**, p. 43-54, 1982.

SVEIBY, K. E. **A nova riqueza das organizações**: gerenciando e avaliando patrimônios de conhecimento. Rio de Janeiro: Campus, 1998.

TAKEUCHI, H.; NONAKA, I. **Gestão do conhecimento**. Porto Alegre: Bookman, 2008.

TURBAN, E.; MCLEAN, E.; WETHERBE, J. **Tecnologia da informação para gestão**: transformando os negócios na economia digital. Porto Alegre: Bookman, 2004.

USP - UNIVERSIDADE DE SÃO PAULO. São Paulo. Tycho : Sistema de apoio à avaliação e a gestão institucional da USP. Disponível em: <<https://uspdigital.usp.br/tycho/apresentacao.jsp?codmnu=1105>>. Acesso em: 28 dez. 2012.

ZHANG, S.; WU, X. Fundamentals of association rules in data mining and knowledge discovery. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, p. 97–116, 2011. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1002/widm.10/full>>. Acesso em: 20 maio 2012.

W3C. Disponível em: <<http://www.w3.org/TR/1998/REC-xml-19980210>>. Acesso em: 26 jan. 2013.

WIJK, J.J. V.; WETERING, H. V. Cushion treemaps: visualization of hierarchical information. **Proceedings of IEEE Information Visualization**, out. p. 73-78. 1999.

YONG, C. S. **Banco de dados**: organização sistemas e administração. São Paulo: Atlas, 1983.

XSREAM. Disponível em: <<http://xstream.codehaus.org/index.html>> Acesso em: 10 jan. 2013.

APÊNDICE A – PADRONIZAÇÃO DO DEPARTAMENTO PARA A MINERAÇÃO DE DADOS

Dado bruto (Departamento conforme cadastrado no Lattes)	Dado Normalizado (Departamento Padronizado)
departamento de quimica depto de quimica departamento academico de quimica e biologia departamento academico de quimica e biologia daqbi departamento de quimica setor de ciencias exatas	departamento de quimica
departamento da teoria e fundamentos da educacao departamento de teoria e fundamentos da educacao departamento de teoria e fundamentos da educacao setor educacao	departamento da teoria e fundamentos da educacao
departameto de teoria e pratica de ensino departamento de teoria e pratica de ensino dtpen departamento de teoria e pratica da educacao departamento de teoria e pratica de ensino	departameto de teoria e pratica de ensino
departamento de fitotecnia e fitossanitarismo depto de fitotecnia e fitossanitarismo	departamento de fitotecnia e fitossanitarismo
departamento de zootecnia; departamento de zootecnia e desenvolvimento rural	departamento de zootecnia
departamento de biologia celular; departamento de biologia; departamento de biologia animal e vegetal	departamento de biologia celular
departamento de bioquimica departamento de bioquimica e biologia molecular depto de bioquimica e biologia molecular	departamento de bioquimica
departamento de farmacologia scb departamento de farmacologia	departamento de farmacologia
departamento de ciencias fisiologicas departamento de fisiologia departamento de fisiologia ufpr	departamento de fisiologia
departamento de zoologia departamento de ecologia e zoologia	departamento de zoologia
departamento de administracao geral e aplicada departamento de administracao geral e aplicada daga	departamento de administracao geral e aplicada
departamento de ciencia e gestao da informacao departamento de ciencias e gestao da informacao departamento de ciencia e gestao da informacao decigi	departamento de ciencia e gestao da informacao
departamento de contabilidade departamento de contabilidade sa	departamento de contabilidade
departamento de informatica departamento de informatica setor de ciencias exatas	departamento de informatica
departamento de matematica departamento de matematica setor de ciencias exatas departamento de matematica instituto de ciencias exatas	departamento de matematica
departamento de artes departamento de artes e musica departamento de artes e letras	departamento de musica artes visuais

Dado bruto (Departamento conforme cadastrado no Lattes)	Dado Normalizado (Departamento Padronizado)
departamento de comunicacao social departamento de comunicacao social departamento de comunicacao	departamento de comunicacao social
departamento de desenho departamento de desenho tecnico departamento academico de desenho industrial departamento de expressao grafica	departamento de expressao grafica
departamento de linguistica letras classicas e vernaculas departamento de linguistica letras classicas e vernaculas	departamento de linguistica letras classicas e vernaculas
departamento de geomatica departamento de geociencias	departamento de geomatica
departamento de arquitetura	departamento de arquitetura urbanismo
departamento de mecanica departamento de engenharia mecanica	departamento de engenharia mecanica
departamento de oceanografia biologica	centro estudos do mar
departamento de eletricidade departamento engenharia eletrica	departamento engenharia eletrica
departamento de educacao departamento de ensino de pos graduacao	nulo
departamento de tecnologia em materiais dptm departamento de tecnologia de materiais dptm departamento de materiais dptm	departamento engenharia civil
departamento de microbiologia e parasitologia ccb	departamento de ciencias biologicas
departamento de engenharia de alimentos	departamento de engenharia quimica
departamento de ciencias agrarias e biologicas	nulo
departamento de silvicultura e manejo	departamento de engenharia e tecnologia florestal
departamento de plantas forrageiras e agrometeorologia	departamento de zootecnia
departamento engenharia ambiental	departamento de engenharia civil

APÊNDICE B – PADRONIZAÇÃO DA FORMAÇÃO PARA A MINERAÇÃO DE DADOS

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
curso de administração	administracao
administracao de empresas	
administracao de empresas	
administração habilitação comércio exterior	
administração de empresas e marketing	
administração	
administração de empresas	
tecnologia em gestão pública	
faculdade de agronomia	agronomia
agronomia	
bacharelado em antropologia	antropologia
antropologia	
arqueologia	arqueologia
arquitetura e urbanismo	arquitetura
arquitetura	
artes cênicas habilitação em direção teatral	artes plasticas
belas artes artes plásticas	
edu artística licenc plena em artes plásticas	
superior de pintura	
artes plásticas	
artes cências habilitação em interpretação teatral	
biblioteconomia	biblioteconomia
biblioteconomia e documentação	
curso de biblioteconomia e documentação	
licenciatura em ciencias biologicas	biologia
ciências biológicas bacharelado	
bacharelado em ciências biológicas	
licenciatura plena em ciências biológicas	
ciências biológicas ecologia	
ciências biológicas licenciatura	
ciências biológicas licenciatura plena	
ciências biológicas - bacharelado	
licenciatura em ciências biológicas	
ciências biológicas licenciatura e bacharelado	
licenciatura plena em ciências biológicas	
bacharelado e licenciatura em ciências biológicas	
ciências biológicas modalidade médica	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
ciências biológicas	
ciências habilitação em biologia	
licenciatura e bacharelado em ciências biológicas	
ciências biológicas (botucatu)	
ciencias biológicas	
ciencias biologicas	
biologia	
ciências biológicas biologia marinha	
licenciatura em ciências biológicas	
ciências biológicas- modalidade médica	
licenciatura e bacharelado em biologia	
bacharelado em biologia	
licenciado em ciências biológicas habilitado em bi	
ciências biológicas	
licenciatura en ciencias biológicas	
licenciatura em ciências biológicas	
bacharelado em ciências biológicas	
botânica aplicada às plantas medicinais turma ii	botanica
processamento de dados	
bacharelado em ciencias da computacao	
curso superior de tecnologia de computação	
bacharelado em ciência da computação	
bacharelado em informática	
informatica	
ciencia da computacao	
ciências da computação	
bacharel em informática	
bacharelado em informática	
análise de sistemas	
tecnólogo em processamento de dados	
ciência da computação	
informática	
bacharelado em ciência da computação	
ciências licenciatura curta	
bachelor of science	
ciências naturais	
ciências licenciatura	
biociências	
licenciatura em ciências	
licenciatura em ciências	
ciências contábeis	ciencias contabeis

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
ciencias contabeis	ciencias contabeis
contabilidade	
licenciatura de sociologia	ciencias sociais
bacharel em ciências sociais	
graduação e licenciatura em ciências sociais	
licenciatura em ciências sociais	
sociologia	
ciências sociais (modalidade: política)	
ciências sociais bacharelado	
estudos sociais	
ciencias sociais	
faculté des sciences politiques et sociales	
bacharelado em sociologia	
ciências sociais habilitação sociologia	
ciências sociais licenciatura	
ciências sociais	
ciências sociais (espec. c. política e sociologia)	
licenciatura em estudos sociais	
departamento de ciências sociais	
licenciatura em ciencias sociais	
ciências sociais/sociologia (bacharelado)	
ciências sociais/sociologia (licenciatura)	
jornalismo	comunicacao social
comunicação social jornalismo	
comunicação social habilitação jornalismo	
comunicação social habilitação relações públicas	
curso de comunicação visual	
comunicacao social	
comunicação	
comunicação social	
comunicação visual	
comunicação social - jornalismo	
desenho industrial projeto do produto	desenho
desenho industrial	
licenciatura em desenho	
desenho - licenciatura em	
direito	direito
bacharelado em direito	
bacharelado em direito	
ciências econômicas e empresariais-seção economia	economia
economia	
ciências econômicos	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
curso de ciências econômicas	economia
economia incompleto (até o 3º ano)	
ciências econômicas	
ciencias econômicas	
licenciatura em educação artística música	educacao artistica
educação artística	
educação artística - habilitação em música	
educação artística habilitação em artes plásticas	
licenciatura plena em educação física	educacao fisica
educação física	
licenciatura em educação física	
bacharelado em educação física	
licenciado em educação física	
licenciatura em educação física	
educação física e técnico desportivo	
enfermagem habilitação em licenciatura	enfermagem
enfermagem e obstetrícia	
enfermagem	
licenciatura em enfermagem	
enfermagem e obstetrícia	
graduação em enfermagem	
engenharia	engenharia
engenharia agrônoma	engenharia agricola
engenharia agrícola	
engenharia agronomica	
engenharia cartografica	engenharia cartografica
engenharia cartográfica	
engenharia cartográfica	
engenharia de operação construção civil	engenharia civil
bacharelado em engenharia civil	
engenharia civil	
tecnologia da construção civil	
engenharia da computação	engenharia da computacao
engenharia de alimentos	engenharia de alimentos
engenharia de alimentos	
engenharia elétrica (eletrônica/telecomunicações)	engenharia eletrica
engenharia industrial elétrica - ênfase em eletrôn	
engenharia elétrica	
engenharia eletrica	
engenharia industrial eletrica	
ingenieria eléctrica	
engenharia industrial eletrônica	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
engenharia radiotécnica	engenharia eletrica
curso de engenharia industrial elétrica ênfase ele	
engenharia eletrônica	
engenharia eétrica	
engenhariav florestal	engenharia florestal
curso de engenharia florestal	
engenaharia florestal	
ingeniería forestal	
engenharia florestal	
engenharia industrial madeireira	engenharia madeireira
engenharia industrial orientação mecânica	engenharia mecanica
filosofia	
tecnologia em mecânica	
filosofia	
filosofia	
engenharia mecânica e automóveis	
engenharia mecanica	
engenharia mecânica ênfase construção de máquinas	
engenharia mecânica	
filosofia	
filosofia	
engenharia metalurgia e dos materiais	engenharia metalurgica
engenharia metalúrgica	
engenharia quimica	engenharia quimica
engenharia química	
graduação em engenharia química	
engenharia química	
estatística	estatistica
bacharelado em estatística	
farmacia-analise clinicas	farmacia
farmácia bioquímica	
farmacia e bioquimica	
farmácia bioquímica e industria	
farmácia habilitação indústria	
farmácia industrial	
farmácia habilitação em análises clínicas	
farmacêutico industrial	
farmácia e análises clínicas	
farmacia	
farmácia	
farmacêutico bioquímico	
curso de farmácia bioquímica	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
farmácia habilitação em indústria	farmacia
farmácia e bioquímica	
física	física
licenciatura em física	
física - bacharelado	
física	
física	
bacharelado em física	
bacharelado em física	
bacharelado em física	
licenciatura em física	
bacharelado em física	
licenciatura em física	
licenciatura plena em física	
faculdade de aerofísica e pesquisa espacial	
fisioterapia	fisioterapia
fonoaudiologia	fonoaudiologia
licenciado e bacharel em geografia	geografia
bacharelado em geografia	
bacharelado em geografia	
geographie studiengang diplom	
licenciatura en ciencias geológicas	
geografia (licenciatura plena)	
geografia (bacharel)	
licenciatura em geografia	
bacharelado em geografia	
licenciatura plena em geografia	
bacharelado em geografia - complementação	
bacharel em geografia	
geografia	
curso de geologia	geologia
geologia	
geologia	
história e geografia	historia
curso de história natural	
geografia e historia - esp. antropologia americana	
licenciatura plena em história	
licenciatura em história	
bacharelado em história	
história - licenciatura	
licenciatura plena em história	
licenciatura em história natural	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
história	historia
historia	
história natural	
departamento de história	
geografia e historia	
historia natural	
graduação em história	
lettres et civilisation étrangères	letras
letras francês	
letras estrangeiras modernas	
letras espanhol inglês	
bacharelado em letras latim	
licenciado em letras vernáculas	
bacharelado em letras	
bacharelado e licenciatura em letras	
licenciatura em letras alemão português	
licenciatura em letras português e inglês	
bacharelado em letras - português e inglês	
letras	
letras português inglês espanhol	
letras português inglês	
letras português	
bacharelado em lingüística	
cours supérieur d'études françaises	
letras licenciatura em português e inglês	
licenciatura em letras	
curso de letras	
letras italiano	
letras alemão-português	
licenciatura em letras português	
letras (língua e literatura alemã)	
bacharelado em lingüística	
bacharelado e licenciatura português inglês	
bacharelado e licenciatura em português	
diplôme supérieur de langue e littérature français	
matematica aplicada e computacional	matematica
licenciado en matematicas mencion cum laude	
bacharelado em matematica	
bacharelado em matemática	
bacharel em matemática	
licenciatura en ciencias con mencion en matematica	
licenciatura em matemática	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
bacharelado em matemática	matematica
licenciatura em matemática	
licenciatura plena em matemática	
matematica	
licenciatura e bacharelado em matemática	
matemática	
applied mathematics	
aprovação para equivalência diploma de médico nos	medicina
medicina	
medicina	
medicina veterinaria campus botucatu	medicina veterinaria
medicina veterinária	
veterinária	
medicina veterinaria	
clínica cirúrgica veterinária	
meteorologia	meteorologia
bacharel em composição	musica
bacharelado e licentura plena em regência	
bacharelado em piano	
licenciatura em música	
bacharelado em composição	
licenciatura em música	
bacharelado em música	
bacharelado em instrumento	
musicoterapia	musicoterapia
nutrição	nutricao
oceanologia	oceanografia
oceanografia	
faculdade de odontologia	odontologia
curso de odontologia	
licenciatura en odontología	
odontologia	
bacteriologia y laboratorio clinico	outros
bacharelado	
estagiaria	
formação de oficiais da marinha do brasil	
reabilitação	
estudos da américa latina	
licenciatura em pedagogia	pedagogia
licenciatura plena, programa especial de formação	
licenciatura em 2o grau	
pedagogia	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
pedagogia	pedagogia
licenciatura plena em pedagogia	
licenciatura em pedagogia	
licenciatura em educação especial	
formação de professores	
pedagogia habilitação supervisão escolar	
esquema i curso superior de formação de professore	
graduação em psicologia	psicologia
bacharelado em psicologia	
psicologia	
bacharel em psicologia	
licenciatura em psicologia	
licenciatura em psicologia	
bacharelado e licenciatura em psicologia	
formação de psicólogos	
licenciatura e formação de psicólogo	química
bacharelado e licenciatura em química	
bacharel em química	
bacharel e licenciatura em química	
licenciatura em ciências químicas	
bacharel em química com atribuições tecnológicas	
licenciatura em química	
licenciatura em ciências habilitação em química	
licenciatura e bacharelado em química	
química	
bacharelado em química com atribuições tecnológica	
licenciatura plena em química	
química - bacharelado, industrial, biotecnológico	
bacharelado em química tecnológica	
bacharelado em química	
química	
ciências química	
química industrial	
química industrial	
química bacharelado	
químico laboratorista	
química bacharelado com atribuições tecnológicas	
bacharelado em química	
licenciatura em química	
bacharelado e licenciatura em química	
bacharel em química	
licenciatura e bacharelado em química	

Dado Bruto (Formação conforme cadastrado no Lattes)	Dado Padronizado (Formação padronizada)
bacharelado em química	química
turismo	turismo
bacharelado em turismo	
zootecnia	zootecnia
zootecnia agrícola	

APÊNDICE C – RESULTADO ALGORITMO C4.5 DAS REGRAS COM INSTÂNCIAS CLASSIFICADAS

SE ÁREA 2	E ÁREA 3	ENTÃO ÁREA 1
Fisiologia	nulo	Zootecnia
Fisiologia	Ecologia	Fisiologia
Fisiologia	Farmacologia	Farmacologia
Fisiologia	Fisiologia	Farmacologia
Fisiologia	Biologia Geral	Bioquímica
Fisiologia	Educação Física	Fisiologia
Fisiologia	Medicina	Educação Física
Ciência e Tecnologia de Alimentos		Engenharia Química
Biologia Geral		Bioquímica
Psicologia	Saúde Coletiva	Psicologia
Psicologia	Nutrição	Fisiologia
Química Industrial		Recursos Florestais e Engenharia Florestal
Botânica	Farmacologia	Química
Botânica	Medicina	Bioquímica
Botânica	Genética	Agronomia
Engenharia de Materiais e Metalúrgica	Física	Engenharia de Materiais e Metalúrgica
Engenharia de Materiais e Metalúrgica	Química	Química
Engenharia de Materiais e Metalúrgica	Engenharia Civil	Química
Engenharia de Materiais e Metalúrgica	Engenharia Mecânica	Engenharia Mecânica
Bioética		Saúde Coletiva
Arquitetura e Urbanismo		História
Biofísica		Bioquímica
Engenharia de Produção		Engenharia Elétrica
Serviço Social		Geociências
Oceanografia		Botânica
Recursos Pesqueiros e Engenharia da Pesca		Zoologia
Desenho Industrial		Engenharia de Produção

SE ÁREA 2	E ÁREA 3	ENTÃO ÁREA 1
Parasitologia		Bioquímica
Planejamento Urbano e Regional		Geografia
Engenharia de Minas		Engenharia Civil
Saúde Coletiva		Farmacologia

ANEXO A – PARECER DO COMITÊ DE ÉTICA EM PESQUISA DA UFPR

UNIVERSIDADE FEDERAL DO
PARANÁ - SETOR DE
CIÊNCIAS DA SAÚDE/ SCS -



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Construção de modelo para mapear o conhecimento científico dos pesquisadores inseridos nos programas de Pós-Graduação da Universidade Federal do Paraná

Pesquisador: Elaine Conceição Venâncio Santos

Área Temática:

Versão: 1

CAAE: 03557312.9.0000.0102

Instituição Proponente: Universidade Federal do Paraná - Setor de Ciências da Saúde/ SCS

DADOS DO PARECER

Número do Parecer: 156.640

Data da Relatoria: 28/11/2012

Apresentação do Projeto:

A pesquisa de investigação qualitativa caracteriza-se pela abordagem descritiva, tendo como foco a gestão do conhecimento, apresentando a criação de um modelo capaz de resgatar informações de pesquisadores, e identifica-los a partir de características de uma determinada população em relação a suas variáveis. Para a realização da pesquisa serão coletados dados existentes na Plataforma Lattes dos professores/pesquisadores vinculados aos programas de pós-graduação stricto sensu da UFPR. Estes dados são utilizados para apresentação do modelo de mapa do conhecimento proposto. Para viabilizar a composição da base de dados deste estudo, serão utilizados arquivos XML da base da

Plataforma Lattes, mantida pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), sendo utilizado os mesmos dados que são apresentados no Currículo Lattes e de acesso público, estes arquivos serão repassados pela Pró-Reitoria de Pesquisa e Pós-graduação da UFPR. O universo deste estudo compreende os docentes da Universidade Federal do Paraná, tendo como amostra, 1200 acessos.

Objetivo da Pesquisa:

Justificando a dificuldade em localizar o conhecimento em instituições de ensino e por consequência o compartilhar do conhecimento existente, relacionada :a) à grande quantidade de dados e informações disponíveis na organização de forma não organizada;b) a problemas relativos à gestão de pessoas, criando barreiras ao fluxo e divulgação da informação;c) à ausência de

Endereço: Rua Padre Camargo, 280

Bairro: 2º andar

CEP: 80.060-240

UF: PR

Município: CURITIBA

Telefone: (41)3360-7259

E-mail: cometica.saude@ufpr.br

UNIVERSIDADE FEDERAL DO
PARANÁ - SETOR DE
CIÊNCIAS DA SAÚDE/ SCS -



ferramentas voltadas para a localização do conhecimento já

registrado, tem por Objetivo Primário desenvolver um modelo para a construção de mapa do conhecimento com o objetivo de localizar e indicar o conhecimento existente em instituições acadêmicas, explorando múltiplas técnicas de visualização da informação.

Avaliação dos Riscos e Benefícios:

Não há riscos observáveis e, como Benefícios, o estudo contribuirá para o compartilhamento do conhecimento na instituição, o que vai ao encontro da política de qualificação e capacitação dos professores/pesquisadores, por apresentar uma forma de compartilhar o conhecimento entre as pessoas que atuam na instituição, o que pode apoiar a qualificação/capacitação das pessoas que atuam na UFPR por promover a identificação dos professores/pesquisadores e dos documentos produzidos por estes. No Termo de Consentimento institucional, o Pró-Reitor de pesquisa enfatiza a importância e benefício para o Setor.

Comentários e Considerações sobre a Pesquisa:

Para a realização da pesquisa serão coletados dados existentes na Plataforma Lattes dos professores/pesquisadores vinculados aos programas de pós-graduação stricto sensu da UFPR. Estes dados são utilizados para apresentação do modelo de mapa do conhecimento proposto. A partir dos dados disponíveis na base são consideradas as informações abaixo: 1) áreas de interesse, que abrange: Formação Acadêmica e Áreas de Atuação; 2) áreas de conhecimento, referente às publicações científicas, sendo composta por: artigos publicados em revistas, capítulos de livros, livros. 3) áreas de aplicação: a fim de descobrir a(s) área(s) de aplicação da pesquisa dos professores/ pesquisadores, é realizado o levantamento das palavras-chave da produção científica cadastradas na base, referente às publicações científicas consideradas no item anterior, sendo consideradas as três primeiras palavras-chave que estiverem cadastradas. Na fase seguinte são aplicadas técnicas de mineração de dados visando a descoberta de relacionamento entre as áreas de pesquisas. Pretende-se assim classificar os professores/ pesquisadores de acordo com os temas da produção científico e não somente baseado em suas áreas de conhecimento originais. Posteriormente, pretende associar os agrupamentos às áreas de conhecimento do sistema de classificação adotado; 4) colaboradores: das produções científicas cadastradas, também são considerados todos os autores da publicação, a fim de identificar as redes de colaboração; 5) dados pessoais: são utilizadas apenas informações referente ao local de trabalho, como: departamento, setor e programa de pós-graduação que os professores/ pesquisadores estão vinculados. O nome do pesquisador é codificado e não são utilizadas informações como telefone e e-mail.

Entre os resultados provenientes da aplicação do modelo proposto, destaca-se:- a criação de uma base de dados para o mapa do conhecimento;- facilidade de navegação entre as informações

Endereço: Rua Padre Camargo, 280

Bairro: 2º andar

CEP: 80.060-240

UF: PR

Município: CURITIBA

Telefone: (41)3360-7259

E-mail: cometica.saude@ufpr.br

UNIVERSIDADE FEDERAL DO
PARANÁ - SETOR DE
CIÊNCIAS DA SAÚDE/ SCS -



cadastradas na base de dados, por meio de uma interface em ambiente web;- possibilidade de filtrar as informações e visualiza-las com diferentes técnicas;- possibilidade de inferir diferentes formas de relacionamento entre os pesquisadores/professores e departamentos.

A pesquisa está bem estruturada, mostra-se relevante e vem sanar uma lacuna no que tange ao banco de dados da produção acadêmica na UFPR.

Considerações sobre os Termos de apresentação obrigatória:

Pede a dispensa deo Termo de Consentimento Livre e Esclarecido e justifica que, não será realizado contato pessoal para obtenção dos dados dos pesquisadores vinculados aos programas de pós-graduação da Universidade Federal do Paraná e que os dados utilizados serão os que constam na Plataforma Lattes obtidos através dos arquivos XML disponibilizados pela Pró-Reitoria de Pesquisa e Pós-Graduação da UFPR. Os dados que estão disponíveis nos arquivos XML da Plataforma Lattes, contem as informações que os usuários do Currículo Lattes cadastram no sistema. O acesso ao arquivo XML facilita a leitura de grandes quantidades de dados de forma automática pelo programa que será desenvolvido para ler os dados. Conforme a CNPq "A Plataforma Lattes é uma base de dados pública, tanto no que se refere ao ingresso quanto à recuperação das informações através da Internet. Apesar da disponibilização pública das informações, instituições de ensino, pesquisa e inovação do País têm solicitado ao CNPq o acesso aos dados curriculares de seus professores, pesquisadores, alunos e colaboradores, com o objetivo integrar os dados

do Lattes aos seus sistemas de informação, gerar indicadores internos de produção científica e tecnológica, realizar estudos através da aplicação de ferramentas de mineração de dados, e apoiar a implementação de políticas de gestão." No endereço:

<http://impl.cnpq.br/impl/Gramaticas/Curriculo/DTD/Documentacao/DTDCurriculo.pdf>, conta exemplo do formato XML.

Os demais termos atendem ao solicitado pelo Comitê de Ética.

Recomendações:

Não há.

Conclusões ou Pendências e Lista de Inadequações:

O projeto atende às normas do Comitê de Ética em Pesquisa

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Endereço: Rua Padre Camargo, 280

Bairro: 2º andar

UF: PR

Município: CURITIBA

Telefone: (41)3360-7259

CEP: 80.060-240

E-mail: cometica.saude@ufpr.br

UNIVERSIDADE FEDERAL DO
PARANÁ - SETOR DE
CIÊNCIAS DA SAÚDE/ SCS -



Considerações Finais a critério do CEP:

Solicitamos que sejam apresentados a este CEP, relatórios semestrais sobre o andamento da pesquisa, bem como informações relativas às modificações do protocolo, cancelamento, encerramento e destino dos conhecimentos obtidos, através da Plataforma Brasil - no modo: NOTIFICAÇÃO.

CURITIBA, 28 de Novembro de 2012

Assinador por:
IDA CRISTINA GUBERT
(Coordenador)

Endereço: Rua Padre Camargo, 280

Bairro: 2º andar

UF: PR

Município: CURITIBA

Telefone: (41)3360-7259

CEP: 80.060-240

E-mail: cometica.saude@ufpr.br

ANEXO B – AUTORIZAÇÃO DA PRPPG PARA COLETA DOS DADOS

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

Curitiba, 6 de novembro de 2012
Ofício 264/2012 – PRPPG/UFPR

**À Comissão de Ética em Pesquisa
Setor de Ciências da Saúde
Nesta Universidade**

Assunto: manifesta interesse institucional

Senhor Coordenador

Declaramos que a Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Paraná está de acordo com a condução do projeto de pesquisa **“Construção de modelo para mapear o conhecimento científico dos pesquisadores inseridos nos programas de Pós-Graduação da Universidade Federal do Paraná”** sob a responsabilidade de Elaine Conceição Venâncio Santos, tão logo o projeto seja aprovado pelo Comitê de Ética em Pesquisa do Setor de Ciências da Saúde da UFPR, e até o seu final em 30 de março de 2013.

Estamos cientes que a fonte da pesquisa serão os dados da Plataforma Lattes referente aos professores dos programas de pós-graduação da UFPR, bem como de que o presente trabalho deve seguir as resoluções atinentes ao projeto em apreço.

Outrossim, informamos que, uma vez cientes do parecer exarado pelo Comitê de Ética, poderemos emitir outro documento em conformidade com exigências adicionais que houver, autorizando a execução do projeto em tela.

Certos da atenção, enviamos nossas cordiais saudações.

Respeitosamente,

Prof. Dr. Sergio Scheer
Pró-Reitor de Pesquisa e Pós-Graduação
UFPR